THE QUEST FOR OSMOTIC STRESS MARKERS IN *MUSA*: FROM PROTEIN TO GENE AND BACK IN A NON-MODEL CROP

Anne-Catherine VANHOVE

Supervisor:
Prof. S. Carpentier, KU Leuven
Co-supervisor:
Prof. R. Swennen, KU Leuven

Members of the Examination Committee:
Prof. D. Springael, Chairman, KU Leuven
Prof. B. Cammue, KU Leuven
Prof. J. Robben, KU Leuven
Prof. E. Waelkens, KU Leuven
Prof. S. Wienkoop, Universität Wien

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor in Bioscience Engineering

July 2014

Doctoraatsproefschrift nr. 1194 aan de faculteit Bio-ingenieurswetenschappen van de KU Leuven

Uitgegeven in eigen beheer, Anne-Catherine Vanhove, Lindelaan 7, 3001 Heverlee

D/2014/11.109/34

Acknowledgements

De afgelopen jaren kon ik op de steun van vele mensen rekenen om mijn doctoraat met als resultaat dit proefschrift tot een goed einde te brengen.

In eerste instantie denk ik dan natuurlijk aan mijn promotor, prof. Sebastien Carpentier. Seb, ik leerde je kennen in oktober 2006 toen je zelf nog doctoraatsstudent was en ik de kans kreeg om het Laboratorium Tropische Plantenteelt te bezoeken. Acht jaar later hebben we een hele weg samen afgelegd en ben ik de eerste doctoraatsstudent die zal afstuderen met jou als promotor. Je eindeloze energie, je enthousiasme en onze (soms urenlange) interessante discussies stuwden me steeds vooruit en droegen in grote mate bij tot dit eindresultaat. Bedankt!

Daarnaast ben ik ook mijn co-promotor prof. Rony Swennen zeer dankbaar voor zijn vertrouwen in mij de afgelopen jaren.

Vervolgens wil ik mijn assessoren, prof. B. Cammue en prof. J. Robben, alsook de andere leden van mijn doctoraatsjury, prof. E. Waelkens, prof. S. Wienkoop en voorzitter Prof. D. Springael, bedanken voor de suggesties die me toelieten dit werk te verbeteren.

Ik dank het agentschap voor Innovatie door Wetenschap en Technologie (IWT) voor de financiële steun die ik als bursaal met de doctoraatsbeurs strategisch basisonderzoek ontving.

Verder wil ik ook de mensen bedanken van de faciliteiten voor massaspectrometrie, SYBIOMA (KU Leuven) en Centre de Recherche Public - Gabriel Lippmann (Luxembourg). Wesley en Ruud, bedankt voor de hulp in het labo op SYBIOMA. Prof. E. Waelkens bedank ik voor het gebruik van de MALDI-TOF/TOF massaspectrometer. Finally, I would like to thank Alberto Cenci (Bioversity International) for his assistance with the annotation of banana genome sequences.

Alle collega's en ex-collega's van het Laboratorium Tropische Plantenteelt ben ik zeer dankbaar voor de hulp, maar zeker ook voor de leuke babbels tijdens de koffie- en lunchpauzes. Een speciaal woord van dank gaat hier uit naar Dr. Bart Panis. Bart, je rust, je advies en je aanmoedigingen gaven me op moeilijkere momenten steeds de energie om door te zetten. Yves en Annick, bedankt voor de hulp in het proteomicslabo, zeker bij het gieten van de gradiëntgels en het scannen. Edwige, bedankt voor de goede zorgen voor mijn plantjes. Lut, bedankt voor de helpende hand waar het kon. Hien, thank you for teaching me qPCR all those years ago. My fellow PhD students (past and present) often provided some much needed relaxation: going to Pukkelpop for a day, watching a World or European Cup match together, getting together for an international meal, going for drinks in the evening, picnicking in the park… To my former PhD colleagues: thank you for your advice and support in the early years. To the current PhD students: good luck!

Uiteraard wil ik ook graag mijn vrienden en familie bedanken. Ook jullie zorgden ervoor dat ik mij buiten de werkuren goed kon ontspannen met o.a. spelletjesavonden, geocache-uitstappen, cursussen Spaans en fotografie, restaurant- en cafébezoekjes, gezellige kookafspraken, bezoekjes, een dagje Brugge en reizen naar Florida, Princeton en Boston. Jullie zijn met te veel om allemaal op te noemen, maar bedankt voor alle leuke momenten! Stijn, de afwas-babbelgelegenheden en de gezellige thee-avonden op het appartement zorgden regelmatig voor de nodige ontspanning na de werkdag. Ik had het niet beter kunnen treffen met mijn huisgenoot. Jo en Leen, jullie wil ik in het bijzonder even bedanken. Jullie deur staat altijd voor mij open en jullie onvoorwaardelijke steun betekent veel voor mij. Een dik anderhalf jaar geleden gaven jullie me een bijzonder mooi cadeau in de vorm van mijn metekindje, Lisa. Ze tovert elke keer weer een lach op mijn gezicht. Bedankt!

Tot slot dank ik ook graag mijn ouders en mijn broer. Dominique, ook jij hebt enorm hard gewerkt de laatste jaren, maar je was er altijd om me te helpen waar het kon. Mama en papa, jullie staan altijd voor me klaar, zeker op de moeilijkere momenten. Jullie steun en jullie vertrouwen in mij hebben me altijd gestimuleerd om mijn dromen na te jagen. Bedankt!

Voor Daphné,
Je hebt een steentje verlegd in mijn rivier op aarde,
het water gaat er voor altijd anders dan voorheen…

Anne-Catherine

Summary

Bananas and plantains are a major staple food and export product in more than 120 countries with a worldwide production of over 135 million tonnes per year. The Laboratory of Tropical Crop Improvement hosts the Bioversity International *Musa* Germplasm Transit Centre which contains the world's largest banana collection with over 1400 accessions kept as *in vitro* plants.

Water is one of the most limiting abiotic stress factors in banana production. We therefore designed a long term experimental set-up to screen the available *Musa* biodiversity for drought tolerance in which osmotic stress research is a first step. This research was executed at three levels: cell cultures, heterotrophic *in vitro* plants and autotrophic plants.

Research on banana cell cultures identified more than fifty potential osmotic stress markers via proteomics and transcriptomics. To evaluate the suitability of these stress markers for future use in high-throughput screening of banana varieties, we assessed the four most promising via qPCR. We showed that all four candidates reacted to the stress treatment. One (phosphoglycerate kinase) was validated as an osmotic stress marker.

Then our focus shifted from the model on cell cultures towards the plant level. We developed a heterotrophic *in vitro* growth model to screen five varieties representing different genome constitutions present in *Musa*. The proteome of the variety with the smallest growth reduction was analyzed by two-dimensional gel electrophoresis. We successfully identified 24 proteins as potential osmotic stress markers of which five (PR10, isoflavone reductase, glutathione-S-transferase, S-adenosyl methionine synthase and phosphoglucomutase) had already been identified in cell cultures and we showed that proteins belonging to the defense and reactive oxygen species metabolism and to the energy metabolism contributed to the new homeostasis in the stressed *in vitro* plants.

Further proteomic research on autotrophic plants again revealed 35 potential stress markers of which six (HSP20, HSP70, glutathione-S-transferase, S-adenosyl methionine synthase, sucrose synthase and phosphoglycerate kinase) had already been identified in cell cultures and/or *in vitro* plants. Finally we focused our research on one interesting osmotic stress marker protein family, HSP70. It is not uncommon to identify several spots on a gel from two-dimensional gel electrophoresis with the same general identification of the gene family. Gene families in banana consist of paralogs, genes related by duplication within a genome, and allelic variants, genes at the same locus of homologous chromosomes. HSP70 was identified in a trail of six spots. With the availability of the *Musa* A and B genomes and the combinatorial use of gel-based and gel-free proteomics techniques, we were able to pinpoint in an ABB variety which paralogs and/or allelic variants were expressed and were present in the spots. We also identified an osmotic stress-responsive HSP70 encoded by the paralog located on chromosome 2.

The nine osmotic stress markers (HSP20, HSP70, PR10, isoflavone reductase, glutathione-S-transferase, S-adenosyl methionine synthase, sucrose synthase, phosphoglucomutase and phosphoglycerate kinase) identified in this research should now be screened in several varieties and validated under real drought conditions. Combining the validated stress marker genes with phenotyping approaches will help in the future to diagnose the severity of stress and finally drought stress tolerance marker will aid us in the identification of drought tolerant varieties and facilitate banana breeding for drought tolerance.

Samenvatting

Met een wereldwijde productie van meer dan 135 miljoen ton per jaar zijn bananen een basisbestanddeel in de voeding en een belangrijk exportproduct in meer dan 120 landen. Het Laboratorium Tropische Plantenteelt herbergt het Bioversity International *Musa* Germplasm Transit Centre dat meer dan 1400 accessies bewaart als een *in vitro* collectie.

Water is een van de meest limiterende abiotische stressfactoren in de banaanproductie. Daarom ontwikkelden we een proefopzet om de beschikbare *Musa* biodiversiteit te evalueren voor droogtetolerantie. Osmotisch stressonderzoek is hierbij een eerste stap en werd uitgevoerd op drie niveaus: celculturen, heterotrofe *in vitro* planten en autotrofe planten.

Door middel van proteomics- en transcriptomicsonderzoek identificeerden we meer dan vijftig potentiële osmotische stressmerkers in celculturen van banaan. We analyseerden de expressie van de vier meest belovende stressmerkers via qPCR om hun geschiktheid voor toekomstig gebruik voor een grootschalige evaluatie van banaanvariëteiten te beoordelen. Alle kandidaten bleken te reageren op de stressbehandeling. Een kandidaat (fosfoglyceraat kinase) werd gevalideerd als een geschikte osmotische stressmerker op celniveau.

In een volgende fase maakten we de stap van een evaluatiemodel op celniveau naar een op plantniveau. We ontwikkelden een heterotroof *in vitro* groeimodel om de osmotische stresstolerantie te beoordelen van vijf variëteiten, die de verschillende genoomconstituties van banaan vertegenwoordigden. We analyseerden het proteoom van de variëteit met de kleinste groeireductie via tweedimensionale gelelektroforese. We identificeerden 24 proteïnes als potentiële osmotische stressmerkers waarvan er vijf (PR10, isoflavon reductase, glutathion-S-transferase, S-adenosyl methionine synthase en fosfoglucomutase) reeds geïdentificeerd werden in de celculturen. Proteïnes betrokken bij het verdedigingsmetabolisme, het

metabolisme van de reactieve zuurstofdeeltjes en het energiemetabolisme droegen bij tot een nieuw evenwicht in de gestresseerde *in vitro* planten.

Verder proteomicsonderzoek op autotrofe planten leidde tot de identificatie van 35 potentiële stressmerkers waarvan er zes (HSP20, HSP70, glutathion-S-transferase, S-adenosyl methionine synthase, sucrose synthase en fosfoglyceraat kinase) reeds geïdentificeerd werden in celculturen en/of *in vitro* planten. Ten slotte concentreerden we het onderzoek op één osmotische stressmerker proteïnefamilie, HSP70. Het is niet ongebruikelijk dat verschillende spots op een gel, gemaakt aan de hand van tweedimensionale gelelektroforese, dezelfde algemene genfamilie-annotatie krijgen. Genfamilies in banaan bestaan uit paralogen, genen gerelateerd via duplicatie binnen een genoom, en allelische varianten, genen op dezelfde locus van homologe chromosomen. HSP70 werd geïdentificeerd in een reeks van zes spots. Nu zowel het A- als het B-genoom van *Musa* beschikbaar waren, gebruikten we een combinatie van gelgebaseerde en gelvrije proteomicstechnieken om in een ABB-variëteit exact vast te stellen welke paralogen en allelische varianten tot expressie werden gebracht en aanwezig waren in de spots. Een HSP70 die geëncodeerd werd door de paraloog gelegen op chromosoom 2, reageerde op osmotische stress.

De negen osmotische stressmerkers (HSP20, HSP70, PR10, isoflavon reductase, glutathion-S-transferase, S-adenosyl methionine synthase, fosfoglucomutase, sucrose synthase en fosfoglyceraat kinase), die we in dit doctoraatsonderzoek identificeerden, moeten nu geëvalueerd worden in verschillende variëteiten en gevalideerd worden onder realistische droogtecondities. De combinatie van gevalideerde stressmerkers met fenotypering zal in de toekomst bijdragen tot het bepalen van de ernst van de stress. Verder zullen droogtestresstolerantiemerkers bijdragen tot de identificatie van droogtetolerante variëteiten en tot het kweken van bananen die droogtetoleranter zijn.

Symbols and abbreviations

Symb.-Abbr.	Description
2D-DIGE	two-dimensional difference gel electrophoresis
2DE	two-dimensional gel electrophoresis
ABA	abscisic acid
ABRE	ABA-responsive element
ATP	adenosine triphosphate
DDA	data-dependent acquisition
DIA	data-independent acquisition
DIGE	difference gel electrophoresis
DRE	drought responsive element
DREBs	DRE-binding proteins
emPAI	exponentially modified protein abundance index
EST	expressed sequence tag
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GC-MS	gas chromatography-mass spectrometry
GWAS	genome-wide association study
HILEP	hydroponic isotope labelling of entire plants
HSE	heat shock element
HSP20	20 kilodalton heat shock protein
HSP70	70 kilodalton heat shock protein
ICAT	isotope-coded affinity tags
ICPL	isotope-coded protein label
iTRAQ	isobaric tags for relative and absolute quantitation
LC	liquid chromatography
MALDI	matrix-assisted laser desorption/ionization
MRM	multiple reaction monitoring
MS	mass spectrometry
MS/MS	tandem mass spectrometry

MudPIT	multi-dimensional protein identification technology
NAD(P)H	nicotinamide adenine dinucleotide (phosphate)
NGS	next generation sequencing
NMR	nuclear magnetic resonance
PAI	protein abundance index
pI	isoelectric point
PR10	pathogenesis-related protein 10
PTM	post-translational modification
qPCR	quantitative polymerase chain reaction
QTL	quantitative trait locus
RNA-seq	RNA sequencing
ROS	reactive oxygen species
RP	reversed phase
RuBisCO	ribulose-1,5-bisphosphate carboxylase/oxygenase
SAGE	serial analysis of gene expression
SCX	strong-cation-exchange
SILAC	stable isotope labeling with amino acids in cell culture
SILIP	stable isotope labeling in planta
SNP	single nucleotide polymorphism
SRM	selected reaction monitoring
SUMO	small ubiquitin-like modifier
TOF	time-of-flight
UTR	untranslated region

Table of contents

I. Introduction

Rationale and outline of thesis

Research at the Laboratory of Tropical Crop Improvement at KU Leuven focuses on safeguarding biodiversity and improving tropical crop production with a focus on bananas and plantains. The laboratory hosts Bioversity's International Transit Centre which contains the *Musa* International Germplasm collection with over 1400 accessions kept as an *in vitro* collection. About 850 of these accessions have been successfully cryopreserved as well.

Bananas and plantains are a major staple food and export product in more than 120 countries with a worldwide production of over 135 million tonnes per year (FAO, 2012). They are giant herbaceous monocotyledonous plants which belong to the *Musa* genus (family *Musaceae*, order *Zingiberales*). Cultivated banana varieties are hybrids of two wild diploid species *Musa acuminata* (genome constitution AA) and *Musa balbisiana* (genome constitution BB). Most cultivated varieties are triploids with either an AAA, AAB or ABB genome constitution. Varieties with an AAB or ABB genome constitution are said to be more drought tolerant and hardy due to the presence of the B genome (Simmonds, 1966; Thomas et al., 1998; Robinson and Sauco, 2010). The commercially exploited varieties are triploids with an AAA genome constitution which are sweet and suitable to immature harvesting, transport and ripening upon arrival. However, this AAA Cavendish group is drought sensitive.

Water is one of the most limiting abiotic stress factors in banana production. Bananas need at least 25 mm of water per week and an annual rainfall of 2000-2500 mm evenly distributed along the year is considered optimal for banana production. When there is no access to irrigation, mild drought conditions are responsible for considerable yield losses. Van Asten et al. (2011) calculated a yield loss of up to 65% when the annual rainfall was below 1100 mm. Moreover, in the humid tropics bananas are threatened by the disease Black Sigatoka, caused by the fungus *Mycosphaerella fijiensis*. Export bananas, all from the Cavendish subgroup, are extremely susceptible and economic losses arise from yield loss, premature yellowing and chemical disease control costs. The cultivation of bananas in drier

areas where the infection rate is much lower, is therefore an alternative but then drought stress problems are possible (Marin et al., 2003; Robinson and Sauco, 2010).

To screen *Musa* biodiversity for drought tolerance, we designed a long term experimental set-up (Figure 1). An osmotic stress model has been used to approximate drought in the first two phases. Osmotic stress, similarly to drought stress, causes water deficit in tissues but osmotic stress can be more tightly controlled and can easily be applied to *in vitro* plants. Osmotic stress, however, simulates water deficit by offering water that is less available to the plant while drought is a true lack of water. In a first phase heterotrophic *in vitro* plants have been screened as they provide a highly controlled system to allow fast screening for osmotic stress. In a second research phase autotrophic in vitro plants were used which no longer receive sugar from the medium for growth but perform photosynthesis in growth chambers and have functional stomata and perform real transpiration. The next phase will analyze actual drought stress on greenhouse plants and in a final validation phase field plants will be used. Furthermore, this experimental set-up includes different plant developmental stages. We combine phenotyping with proteomics research in this screening to gain an understanding of the osmotic and drought tolerance mechanisms in banana.

Until 2012 *Musa* was an unsequenced non-model crop. A comparison between proteomics and transcriptomics techniques showed that a proteomics approach at that time offered better characterization for unsequenced crops (Carpentier et al., 2008). The combination of the available EST databases and cross-species identification resulted in significantly more protein identifications in proteomics. Previous proteomics research focused on acclimation research of meristematic cells in relation to cryopreservation and resulted in the identification of more than fifty stress markers.

In this thesis, stress markers are defined as genes with a statistically significant differential abundance at the mRNA level and/or at the protein level in response to the application of a stress or combination of stresses. Osmotic stress markers are defined as genes with a statistically significant differential abundance at the mRNA level and/or at the protein level in response to osmotic stress. Osmotic stress tolerance markers are genes with a differential expression at the mRNA level and/or protein level that can be statistically correlated to osmotic stress tolerance.

In this thesis, we focused first on the validation of potential stress markers from the cell model using qPCR. Subsequently, our focus shifted towards osmotic stress research in plants to identify osmotic stress markers. As the *Musa* reference A

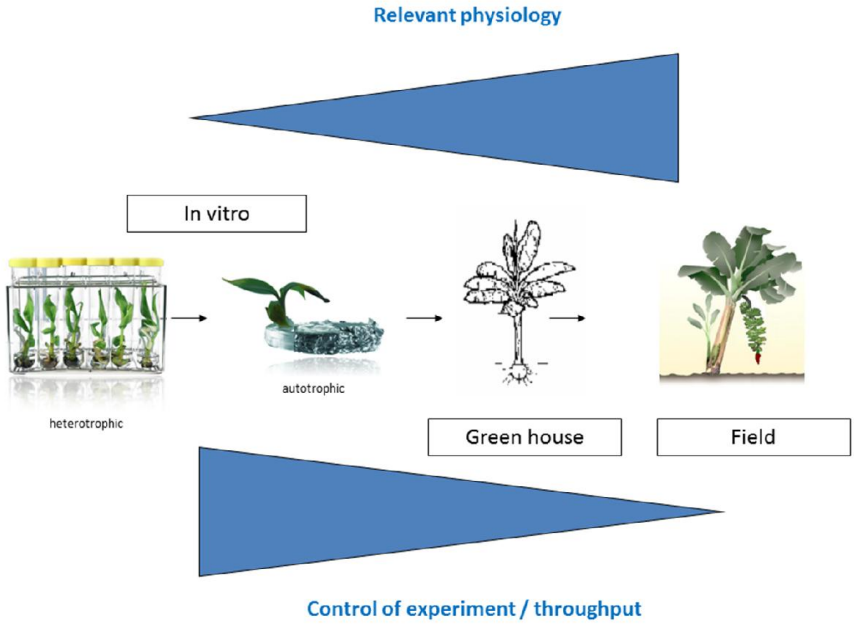

Figure 1: Experimental overview for screening for osmotic/drought tolerance (figure from Vanhove et al. (2012)). In vitro screening (left) deals with osmotic tolerance but offers the advantage that many plants can be screened under controlled conditions. Screening of greenhouse and field plants (right) deals with drought screening under conditions closer to farming conditions but less plants can be analyzed and not all parameters can be controlled.

genome became available in 2012 (D'Hont et al., 2012) followed by a draft B genome in 2013 (Davey et al., 2013), other research possibilities present themselves, yet proteomics research remains important.

In chapter 1 we discuss the available omics approaches and their limitations in crops. We discuss both recent and more established technologies to provide a wide overview of the possibilities in future crop research. Genomics, transcriptomics, proteomics, metabolomics and phenomics approaches are reviewed. In chapter 2, we first discuss drought responses and osmotic stress. We then focus on recent accomplishments in crops abiotic stress research and more specifically on what has been achieved in banana using the omics approaches discussed in the previous chapter.

Chapter 3 deals with the verification of stress marker genes in meristem cultures subjected to osmotic stress. Pathogenesis-related protein 10, SUMO-conjugating enzyme, ABA-responsive protein and phosphoglycerate kinase were identified as potential stress markers in previous proteomics and transcriptomics studies. To

evaluate their suitability for future use in high-throughput screening of varieties, we used qPCR to follow the transcription levels of these potential stress marker genes over time. We showed that all four candidates reacted to the stress treatment and phosphoglycerate kinase was identified as an osmotic stress marker.

In chapter 4 we screened five *Musa* varieties representing the different genome constitutions (AAA, AAAh, AAB, AABp and ABB) using a heterotrophic *in vitro* growth model (Figure 1). The ABB variety showed the smallest growth reduction and was analyzed by two-dimensional difference gel electrophoresis (2D-DIGE) to investigate the new homeostasis. We used two-dimensional gel electrophoresis (2DE), the recommended proteomics technique in unsequenced plants. This approach separates and quantifies whole proteins on gels. The separated proteins are then digested into peptides for identification with mass spectrometry. This approach separates quantification and identification. Gel-free approaches where the whole protein sample is digested into peptides before quantification and identification result in more complex peptide mixtures and it is challenging to predict which peptides belong to the same protein which hinders protein quantification. We successfully identified 24 differential proteins and showed that proteins belonging to the defense and reactive oxygen species metabolism and to the energy metabolism contributed to the new homeostasis in the *in vitro* plants.

In chapter 5 we focused on one particular interesting protein family, HSP70. It is not uncommon to identify several spots on a gel with the same general identification of the gene family. Here, we chose to focus on a trail of 6 spots which were all identified as HSP70 in the osmotic stress model. With the availability of *Musa* A and B genomes and the combinatorial use of gel-based and gel-free approaches, we were able to pinpoint which paralogs and allelic variants were present in the samples. We also identified an osmotic stress-responsive HSP70 isoform located on chromosome 2.

Based on the acquired insights, we place the experimental chapters in perspective in the last chapter. As *Musa* is now a sequenced non-model crop, some opportunities are now available which were not at the start of this project. Consequently, we suggest two workflows for future research in non-model crops. The first one focuses on crop research in general while the second workflow suggests the steps to follow to set-up successful proteomics experiments in sequenced model plants and sequenced and unsequenced non-model crops. Finally, we suggest future research avenues for osmotic and drought stress research in *Musa*.

Chapter 1

Omics approaches and their challenges in non-model crops

1.1 Introduction

The application of omics approaches to crops is not always straightforward. This chapter provides an overview of the available approaches with a focus on research in crops. The last years have brought many new techniques which start to be implemented in crops. This review gives a general overview of established and new omics approaches available to crops, both sequenced and unsequenced. In combination with the research presented further in this dissertation, this will allow us to critically reflect on the performed research as well as to propose a workflow for stress research in crops in the future.

1.2 Genomics

These last years, crop research has been significantly changed with the introduction of fully sequenced genomes for several crops. A reference genome facilitates comparative genomic approaches and provides essential sequence information to transcriptomics and proteomics approaches.

1.2.1 Genome sequencing

In 2013, the cape of 50 published sequenced plant genomes was reached (Michael and Jackson, 2013). By the end of May 2014 more than 70 sequenced plant genomes have been published and additional, as of yet unpublished, plant genomes are already available at databases such as Phytozome[1]. It all started with the sequencing of the model plant *Arabidopsis thaliana* in 2000 (Arabidopsis Genome Initiative, 2000). Later, draft genomes of both the *japonica* and *indica* rice varieties were published and a completed genome sequence of *japonica* rice was released by the International Rice Genome Sequencing Project in 2005 (Goff et al., 2002; Yu et al., 2002; International Rice Genome Sequencing Project, 2005) (Table 1.1). The number of sequenced genomes of other crops steadily increased each year between 2006, when one genome was published, and 2010, which saw the release of 5 new genomes. By 2011 the number of sequenced plant genomes increased dramatically with more than ten publications per year (Table 1.1). At this rate, the next years should bring many more sequenced plant genomes (Michael and Jackson, 2013).

[1] http://phytozome.jgi.doe.gov/

Table 1.1: Published plant genomes (table from Michael & Jackson, July 2013)

	Scientific name	Common name	Year	Type	Division or monocot/dicot	Chr (#)	Size (Mb)	Assembled (Mb)	Assembled (%)	Gene (#)	Repeat (%)	Seq tech	Journal	*PMID*
1	*Arabidopsis thaliana*	arabidopsis	2000	model	dicot	5	125	115	92	25498	14	Sa	Nature	11130711
2	*Oryza sativa*	rice	2002	crop	monocot	12	430	362	84	59855	26	Sa	Science	11935017
3	*Oryza sativa*	rice	2002	crop	monocot	12	420	389	93	61668	NA	Sa	Science	11935018
4	*Oryza sativa*	rice	2005	crop	monocot	12	389	371	95	37544	26	Sa	Nature	16100779
5	*Populus trichocarpa*	black cottonwood	2006	crop	dicot	19	485	410	84	45555	NA	Sa	Science	16973872
6	*Vitis vinifera*	grape	2007	crop	dicot	19	475	487	103	30434	41	Sa	Nature	17721507
7	*Physcomitrella patens*	moss	2008	model	bryophyta	27	510	480	94	35938	16	Sa	Science	18079367
8	*Vitis vinifera*	grape	2007	crop	dicot	19	505	477	95	29585	27	Sa,4	PlosOne	18094749
9	*Carica papaya*	papaya	2008	crop	dicot	9	372	370	99	28629	43	Sa	Nature	18432245
10	*Lotus japonicus*	lotus	2008	model	dicot	6	472	315	67	30799	56	Sa	DNA Research	18511435
11	*Sorghum bicolor*	sorghum	2009	crop	monocot	10	818	739	90	34496	62	Sa	Nature	19189423
12	*Cucumis sativus*	cucumber	2009	crop	dicot	7	367	244	66	26682	24	Sa,I	Nature Genetics	19881527
13	*Zea mays*	maize	2009	crop	monocot	10	2300	2048	89	32540	85	Sa	Science	19965430
14	*Glycine max*	soybean	2010	crop	dicot	20	1115	973	87	46430	57	Sa	Nature	20075913
15	*Brachypodium distachyon*	brachypodium	2010	model	monocot	5	272	272	100	25532	21	Sa	Nature	20148030
16	*Ricinus communis*	castor bean	2010	crop	dicot	10	320	326	102	31237	50	Sa	Nature Biotechnology	20729833
17	*Malus x domestica*	apple	2010	crop	dicot	17	742	604	81	57386	67	Sa,4	Nature Genetics	20802477
18	*Jatropha curcas*	jatropha	2010	crop	dicot	NA	380	286	75	40929	37	Sa,	DNA Research	21149391
19	*Theobroma cacao*	cocoa	2011	crop	dicot	10	430	327	76	28798	24	Sa,4,I	Nature Genetics	21186351
20	*Fragaria vesca*	strawberry	2011	crop	dicot	7	240	210	87	34809	23	4,S,I	Nature Genetics	21186353

Table 1.1 continued

	Scientific name	Common name	Year	Type	Division or monocot/dicot	Chr (#)	Size (Mb)	Assembled (Mb)	Assembled (%)	Gene (#)	Repeat (%)	Seq tech	Journal	PMID
21	*Arabidopsis lyrata*	lyrata	2011	model	dicot	8	207	207	100	32670	30	Sa	Nature Genetics	21478890
22	*Selaginella moellendorffii*	spikemoss	2011	non-model	lycopod	NA	110	213	193	22285	38	Sa	Science	21551031
23	*Phoenix dactylifera*	date palm	2011	crop	monocot	18	658	381	58	28890	40	I	Nature Biotechnology	21623354
24	*Solanum tuberosum*	potato	2011	crop	dicot	12	844	727	86	39031	62	Sa,4,I	Nature	21743474
25	*Thellungiella parvula*	thellungiella	2011	model	dicot	7	140	137	98	30419	8	4,I	Nature Genetics	21822265
26	*Cucumis sativus*	cucumber	2011	crop	dicot	7	367	323	88	26587	NA	Sa,4	PlosOne	21829493
27	*Brassica rapa*	chinese cabbage	2011	crop	dicot	10	485	284	59	41174	40	I	Nature Genetics	21873998
28	*Cannabis sativa*	hemp	2011	crop	dicot	?	820	787	96	30074	NA	4,I	Genome Biology	22014239
29	*Cajanus cajan*	pigeon pea	2011	crop	dicot	11	833	605	72	48680	52	Sa,I	Nature Biotechnology	22057054
30	*Mediucago truncatula*	medicago	2011	model	dicot	8	454	262	58	62388	31	Sa,4,I	Nature	22089132
31	*Setaria italica*	setaria	2012	model	monocot	9	490	423	86	38801	46	I	Nature Biotechnology	22580950
32	*Setaria italica*	setaria	2012	model	monocot	9	510	397	80	35471	40	Sa	Nature Biotechnology	22580951
33	*Solanum lycopersicum*	tomato	2012	crop	dicot	12	900	760	84	34727	63	Sa,4,S,I	Nature	22660326
34	*Cucumis melo*	melon	2012	crop	dicot	12	450	375	83	27427	NA	Sa,4,I	PNAS	22753475
35	*Linum usitatissimum*	flax	2012	crop	dicot	15	373	318	85	43484	24	I	Plant Journal	22757964
36	*Musa acuminata malaccensis*	banana	2012	crop	monocot	11	523	472	90	36542	44	Sa,4,I	Nature	22801500
37	*Gossypium raimondii*	cotton D	2012	crop	dicot	13	880	775	88	40976	60	I	Nature Genetics	22922876

Table 1.1 continued

	Scientific name	Common name	Year	Type	Division or monocot/dicot	Chr (#)	Size (Mb)	Assembled (Mb)	Assembled (%)	Gene (#)	Repeat (%)	Seq tech	Journal	PMID
38	*Azadirachta indica*	neem	2012	crop	dicot	NA	364	NA	NA	20169	13	4,I	BMC Genomics	22958331
39	*Hordeum vulgare*	barly	2012	crop	monocot	7	5100	4980	98	30400	84	NA	Nature	23075845
40	*Pyrus bretschneideri*	pear	2013	crop	dicot	17	527	512	97	42812	53	I	Genome Research	23149293
41	*Citrullus lanatus*	watermelon	2012	crop	dicot	11	425	354	83	23440	45	I	Nature Genetics	23179023
42	*Triticum aestivum*	wheat	2012	crop	monocot	21	17000	3800	22	94000	80	4	Nature	23192148
43	*Gossypium raimondii*	cotton D	2012	crop	dicot	13	880	738	84	37505	61	Sa,4,I	Nature	23257886
44	*Prunus mume*	chinese plum	2012	crop	dicot	8	280	237	85	31390	45	I	Nature Communications	23271652
45	*Cicer arietinum*	chickpea	2013	crop	dicot	8	738	532	72	28269	49	Sa,I	Nature Biotechnology	23354103
46	*Hevea brasiliensis*	rubber tree	2013	crop	dicot	18	2150	1119	52	68955	72	4,S,I	BMC Genomics	23375136
47	*Phyllostachys heterocycla*	moso bamboo	2013	non-model	monocot	24	2075	2051	99	31987	59	I	Nature Genetics	23435089
48	*Oryza brachyantha*	rice relative	2013	non-model	monocot	12	300	263	88	32038	29	I	Nature Communications	23481403
49	*Prunus persica*	peach	2013	crop	dicot	8	265	227	86	27852	37	Sa	Nature Genetics	23525075
50	*Aegilops tauschii*	wheat DD	2013	crop	monocot	7	4360	4244	97	43150	66	4,I	Nature	23535592
51	*Triticum urartu*	wheat AA	2013	crop	monocot	7	4940	4660	94	34879	67	I	Nature	23535596
52	*Nelumbo nucifera*	ancient lotus	2013	non-model	dicot	8	929	804	87	26685	57	I	Genome Biology	23663246
53	*Utricularia gibba*	bladderwort	2013	non-model	dicot	16	77	82	106	28500	3	4,I	Nature	23665961
54	*Picea abies*	norway spruce	2013	crop	gymnosperm	12	19600	12000	61	28354	NA		Nature	23698360
55	*Capsella rubell*	Capsella	2013	Non-model	Dicot	8	219	135	62	26521	NA	Sa	Nature Genetics	23749190

Abbreviations: Sa, Sanger; 4, Roche/454; S, SOLiD; I, Illumina; NA, not reported in primary publication; kb, kilobases; Mb, megabases; Chr, chromosome; PMID, PubMed ID

The main driving force behind this boom in genome sequencing is the lower cost and the higher throughput of the current sequencing technologies. Especially since the introduction of second generation or next generation sequencing technology in sequencing centers in 2008, the price of sequencing has dropped several orders of magnitude from over $5,000 per Mb DNA in 2001 to slightly more than $100 in 2008 and to $0.045 per Mb DNA in 2014 (Wetterstrand, 2014).

1.2.1.1 Sequencing technologies

The first published plant genomes were sequenced with Sanger technology which produces high-quality reads of up to 1000 bp in length. This electrophoretic sequencing technique however requires a cloning step and is therefore costly and low-throughput (Michael and Jackson, 2013). The next generation sequencing techniques use massive parallel sequencing strategies, have a higher throughput and are less costly but produce much shorter sequences (Morey et al., 2013). Deeper coverage is also needed to maintain high accuracy. At the moment, 454 (Roche Diagnostics[2]) and Illumina[3] are the main sequencing platforms used in plant genome sequencing. 454 produces 400-500 bp reads and Illumina reads are only up to a 150 bp. Third generation sequencing technologies show great promise for the future as they will provide longer reads without any need for clonal amplification prior to sequencing which minimizes the introduction of artefacts from earlier PCR cycles (Morey et al., 2013). Single-molecule real-time sequencing developed by Pacific Biosciences[4] uses a sequencing chip which allows 75,000 reactions to be run in parallel with read lengths from 1000 bp to 10 kb. The GridION system and USB-sized MinION nanopore sequencing device were recently introduced by Oxford Nanopore Technologies[5]. Several other third generation platforms are still under development or in the concept stages. Morey et al. (2013) provide a recent, more in-depth review of past, present and future techniques in DNA sequencing.

1.2.1.2 Genome assembly

The mapping and assembly of the reads remain a major bottleneck for the analysis of plant genomes. Large complex plant genomes are particularly challenging for *de novo* assembly (Schatz et al., 2012). Plant genome sizes vary wildly and larger genomes can be quite complex. While the smallest plant genome from the carnivorous plant *Genlisea margaretae* is only 63 Mb, the largest known genome is that of the rare *Paris japonica* at almost 150,000 Mb (Pellicer et al., 2010). The largest

[2] www.454.com

[3] www.illumina.com

[4] www.pacificbiosciences.com

[5] www.nanoporetech.com

sequenced plant genome at the moment is the Norwegian spruce at 19,600 Mb while the overall median size of sequenced plant genomes is approximately 480 Mb. Plant genomes have a repetitive nature mainly due to the accumulation of transposable elements. Repetitive sequences make up anywhere between 3% and 85% of a genome with a median of 43% (Michael and Jackson, 2013). Larger genomes do not have a proportionally larger amount of genes but accumulate more of those repetitive DNA sequences which hinders the assembly (Schnable et al., 2009; Pellicer et al., 2010; Barabaschi et al., 2012; Michael and Jackson, 2013). Aside from repetitive DNA sequences, polyploidy, observed in almost all flowering plant species, further complicates genome assembly (Soltis et al., 2004). Polyploidy leads directly to increased chromosome numbers and DNA content while subsequent DNA losses and rearrangements lead to a pattern of duplication segments across all chromosomes (De Langhe et al., 2010; Jackson et al., 2011). Whole genome duplications therefore lead to large gene families of homologues as was for instance observed in soybean with 75% of the genes present in multiple copies (Schmutz et al., 2010). The assembly is even further complicated due to heterozygosity and there is a need to develop a plant-specific assembler which can take into account both ploidy and heterozygosity (Figure 1.1) (Schatz et al., 2012). The use of a double-haploid variety, which is homozygous, instead of a diploid variety can circumvent this problem as was done in *Musa* (D'Hont et al., 2012).

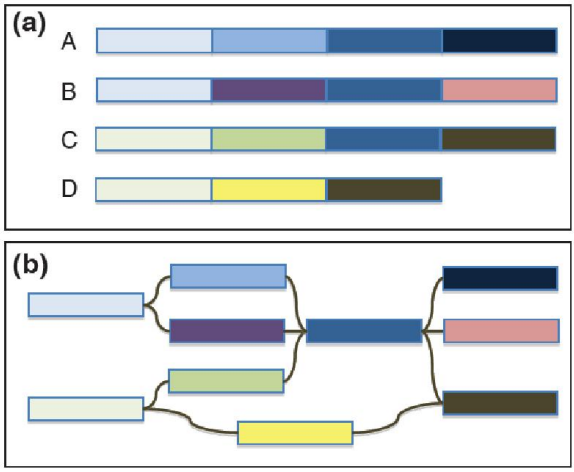

Figure 1.1: Ploidy and heterozygosity complicate genome assembly (figure from Schatz et al., 2012). (a) Schematic representation of a part of a tetraploid genome, consisting of chromosomes A to D with homozygosity/heterozygosity shown as different colored blocks. (b) The assembly graph of the homozygous and heterozygous segments of the genome branch and intertwine in complex patterns. An assembler would need to recognize these branching patterns and attempt to reconstruct the individual sequences for chromosomes A to D.

All of these factors complicate *de novo* genome assembly from short reads as genes from almost identical gene families may be assembled into mosaic gene sequences and the long repetitive stretches generate ambiguity as to how these reads need to be ordered (Schatz et al., 2012). Paired-end sequencing partially alleviates this problem in current genome sequencing projects, but this is probably not sufficient for the longer and more complex repeats found in the larger plant genomes. Longer reads from third generation sequencing technologies promise further improvements and will probably result in a hybrid approach combining short and long reads in the future (Schatz et al., 2012). For now, many genomes still have unresolved nucleotides as there are only a handful of truly finished genomes (Schnable et al., 2009; D'Hont et al., 2012; Schatz et al., 2012; Davey et al., 2013; Michael and Jackson, 2013).

1.2.1.3 Genome annotation

After assembling the reads into a genome sequence, the next step is the annotation, both structural and functional. Structural annotation is composed of two steps: the computational phase followed by the actual annotation phase (Yandell and Ence, 2012). Although several other structural elements, such as non-coding RNAs and transposable elements, contain invaluable information, current annotation pipelines are focused on protein-coding genes since these are the most straightforward to annotate (Yandell and Ence, 2012; Ragupathy et al., 2013). The first step in computational annotation is the masking of the repeat regions. Masking is crucial as failure in this step could lead to extra non-existing genes to be detected downstream or several transposons could be added to existing genes as additional exons. The actual gene prediction these days is mostly based on evidence-driven gene prediction. EST, RNA-seq and protein data are aligned to the assembled genome after which gene prediction tools can use the external evidence to improve their predictions. The gene predictors identify the most likely coding sequence and based on the external data alternative splice isoforms and untranslated regions (UTRs) can be added. Several gene predictors are run and the final annotation is performed by a tool that combines the results of all the gene finders and then selects the best prediction. Yandell and Ence (2012) have written an excellent beginner's guide to eukaryotic genome annotation. Even with these annotation tools, an accuracy of more than 80% is rarely achieved on model organisms, meaning that most gene annotations contain errors (Holt and Yandell, 2011; Yandell and Ence, 2012). These misannotations have a major impact on genetic variation experiments, transcriptomics and proteomics experiments in the same organism, but also on other projects which use these data to structurally annotate their own genomes (Reese and Guigo, 2006; Yandell and Ence, 2012). The final important step in the annotation phase is the functional annotation. This involves homology searches against

databases with genes and proteins with known functions. But not all genes will show enough similarity to be annotated and other genes might be completely new. Furthermore, a general gene annotation does not provide sufficient information as to the function and expression of a gene during different conditions. This is where other omics technologies, such as transcriptomics, proteomics and metabolomics, will provide further information as to how these predicted genes influence the phenotype (Weckwerth, 2011).

1.2.2 Re-sequencing for GWAS, QTL mapping and comparative genomics

Most phenotypes or quantitative traits are the result of several genes and their interaction with the environment. Quantitative trait locus (QTL) analysis is the statistical analysis of both genotypic and phenotypic data to determine the QTL locations, regions in the genome which are at the basis of phenotypic variation for a certain quantitative trait. Both population-based genome-wide association studies (GWAS) and family-based QTL mapping approaches are now used in QTL analysis and will lead to marker-assisted breeding and selection approaches. The population-based GWAS uses a population of unrelated individuals whereas family-based QTL mapping is applied to individuals resulting from several crosses among different founding genotypes (Mitchell-Olds, 2010). With the introduction of the next-generation sequencing technologies, single-nucleotide polymorphisms (SNP) are now increasingly used to link genotypes to phenotypes. Agarwal and colleagues provide an overview of the traditional molecular markers used in plant research including restriction fragment length polymorphism, random amplified polymorphic DNA, amplified fragment length polymorphisms, simple sequence repeats and SNPs (Agarwal et al., 2008). These traditional markers are still extensively used today, but newer alternative markers, mainly gene-targeted functional markers, have been developed as reviewed by Poczai et al. (2013). The population-based GWAS has the advantage of a higher resolution than the traditionally used family-based QTL-mapping which is only based on recombination over a few generations (Mitchell-Olds, 2010). The historical population structure, however, must be taken into account in GWAS as SNP genotypes between different lineages can be neutral or phenotypically important and false positives or false negatives can be identified. This is usually prevented by focusing on a single historical population or lineage. The first step in this whole process is acquiring the phenotypic and SNP data, performing the GWAS or QTL mapping studies and this then results in marker-assisted selection or breeding. While SNP data can now be obtained relatively easily, the phenotyping of large amounts of plants will probably be the main bottleneck in the future. Phenotyping and phenomics is further discussed in section 1.6.

1.2.2.1 Whole-genome re-sequencing

While the genome of many plant species still needs to be sequenced, the whole-genome resequencing of many other plant genomes has already begun. The sequence data generated are aligned to the reference genome and genotypes can be compared to find sequence variants, mutations and structural rearrangements (Varshney et al., 2009). The 1001 genome project is aimed at sequencing 1,001 *Arabidopsis thaliana* accessions from a range of geographic locations. Sequencing finished in 2013 with the last of the genomes expected to be released soon. All sequenced accessions are available as inbred lines for future association studies to identify genetic variation linked to adaptation (1001genomes.org). In a study from 2011 they have already shown that polymorphisms identified in the 80 already sequenced accessions can be used for imputation of polymorphisms in strains that have only been analyzed with a 216k SNP array (Cao et al., 2011). In maize, resequencing of 6 inbred lines, revealed the existence of several presence/absence variations of entire genes between the lines. Their results suggested that gene content complementarity might play an important role in heterosis in maize. They also uncovered about 300 putative genes that are missing in the current maize genome release by resequencing this line (Lai et al., 2010). Other projects have involved sequencing and/or resequencing of soybean, millet and sorghum to identify SNPs between different genotypes (Lam et al., 2010; Zhang et al., 2012; Bekele et al., 2013). In rice a GWAS was performed for 14 agronomic traits in 517 landraces and the identified loci explained on average 36% of the phenotypic variation (Huang et al., 2010). For large and complex genomes, resequencing whole genomes remains difficult. Genotyping-by-sequencing has been specifically developed for those crops to capture a reduced representation of the genome using restriction enzymes. Genotyping-by-sequencing was successfully used in barley and wheat resulting in the identification of 34,000 and 20,000 SNPs respectively (Poland et al., 2012a; Poland et al., 2012b).

1.2.2.2 Targeted re-sequencing

Aside from whole-genome resequencing, targeted (re-)sequencing of several genes to whole exomes is now increasingly used in several plants, both sequenced and unsequenced, using sequence capture approaches. Initially micro-arrays with probes were used to capture the target DNA sequences. This approach was successfully used in maize, which already had a reference genome, to resequence 43 genes in a sequenced and an unsequenced variety (Fu et al., 2010). First a subtraction array was used to discard the repetitive sequences which are overrepresented in plant genomes. In the next step the sequences of interest were captured on a second array and subsequently sequenced. The array-based approach was further modified and resulted in a solution-based hybridization in which biotin-labeled probes are

captured using streptavidin magnetic beads. This approach was used on two unsequenced plants to perform exome capture which selectively captures the coding regions of a genome. In sugarcane, two varieties were compared using probes mainly designed on predicted coding sequences from sorghum, a close relative, as well as some sugarcane ESTs, to enrich for coding sequences in an effort to study SNPs (Bundock et al., 2012). Transcriptome data were used to design probes in pine, an unsequenced species with a particularly large genome of 21.7 Gbp (Neves et al., 2013). While the unknown positions of the introns led to a lower capture efficiency of the exons, they were still able to efficiently enrich and sequence genic portions of two pine species. Avoiding the costly whole genome (re-)sequencing, exome capture can provide a first characterization.

1.2.3 The future for genomics in crops

The NGS techniques have produced a surge in *de novo* genome assemblies, whole-genome re-sequencing and targeted (re-)sequencing at levels from a couple of genes to exomes. These techniques are certainly no longer limited to just model species and have already proven their worth in the analysis of many crops. The size of crop genomes, their repetitive nature, their polyploidy and heterozygosity however still pose problems in the assembly of genomes. Hybrid approaches using the short reads from NGS platforms and longer reads from third-generation sequencing technologies might alleviate this problem in the future. The next couple of years will certainly bring growing numbers of sequenced crops. One such initiative is the recent African Orphan Crops Consortium with an objective of sequencing, assembling and annotating the genomes of 100 traditional African crops[6]. Re-sequencing will provide more information on chromosome structure variation and rearrangements as well as on SNPs, presence-absence variations and novel genes. All these data will provide additional information on annotation of the genes, improve our fundamental understanding of plants and aid in breeding using GWAS, QTL mapping and later marker-assisted breeding or selection. Other omics approaches such as transcriptomics, proteomics, metabolomics and phenomics will however play a significant role into translating this static genome knowledge into dynamic plant response understanding.

[6] http://news.ucdavis.edu/search/news_detail.lasso?id=10804

1.3 Transcriptomics

Probably the easiest way to study changes in gene expression on a genome-wide scale is through transcriptomics. The structure of RNA is homogenous and simple and therefore the analysis is the most straightforward compared to protein and metabolite analysis.

1.3.1 Transcriptome technologies

1.3.1.1 Tag-based approaches

Serial analysis of gene expression (SAGE), developed by Velculescu et al. in 1995, is based on the generation of 15 bp tags from a defined position in each transcript, which are then concatenated, cloned into a plasmid vector and ultimately sequenced using Sanger sequencing (Velculescu et al., 1995). Massively parallel signature sequencing is a more advanced technique based on sequencing of tags. It generates 17 bp tags that are sequenced using a fluorescence-based signature sequencing method on microbeads (Brenner et al., 2000; Reinartz et al., 2002). To analyze the abundance of a transcript, one simply calculates the number of times that a certain tag was found. Though no sequence data needs to be identified *a priori*, the tag needs to be identified as belonging to a gene to convey its biological meaning and this step can be difficult in plants with limited genetic resources. DNA sequences are not as well conserved as amino acid sequences and therefore a cross-species identification based on a short tag is problematic. The development of superSAGE in which longer tags were generated (26 bp) (Matsumura et al., 2003), made the SAGE-approach feasible for unsequenced non-model organisms though still challenging (Coemans et al., 2005; Carpentier et al., 2008).

1.3.1.2 Microarrays

At the same time microarrays, which were significantly cheaper and more high-throughput, were also developed for transcriptomics studies. Microarrays use known probes that will hybridize with the labeled sample and based on the intensity of these dyes transcript levels are estimated. This however implies that sequence information exists before generation of the microarray and this is seriously limited in non-model crops. The limited sequence availability in non-models can be overcome by the use of microarrays of closely related species or by the generation of a species-specific microarray based on known EST data for instance, but these analyses will be less informative (Davey et al., 2009; Pariset et al., 2009). Since more and more sequence and annotation information becomes available from NGS technologies, microarray analysis can be used in an increasing number of species although other limitations remain. Microarray analysis is hampered by high

background noise due to cross-hybridization as well as saturation of signals. Microarrays therefore have a limited sensitivity and dynamic range. Furthermore, microarrays are closed platforms as unknown transcripts cannot be detected. Whole-genome tiling arrays may provide part of the solution towards the detection of new gene transcripts in a sequenced organism but are more expensive and still suffer from the general drawbacks of a microarray approach (Valdés et al., 2013).

1.3.1.3 RNA-seq

With the availability of the next generation sequencing technology, also came the possibility to sequence mRNA, or rather mRNA converted into cDNA, in a high-throughput way, greatly reducing costs. This technique, termed RNA-seq, has clear advantages for non-model and model organisms over the other transcriptomics methods: no previous sequence knowledge is required and a higher sensitivity and dynamic range can be achieved (Wang et al., 2009). After fragmentation of the mRNA and cDNA synthesis, adaptors are fitted to the fragments and sequencing is performed on one of the NGS platforms. Reads can be aligned to a known reference genome but *de novo* assembly of reads into contigs is also possible which increases the use of this technique for crops whose whole genome is not sequenced as was demonstrated in e.g. wheat, agave and horse gram (Bhardwaj et al., 2013; Gross et al., 2013; Oono et al., 2013b). When a reference genome is already available, RNA-seq can provide additional information necessary to identify previously unknown gene coding sequences. Furthermore the data can also be used to improve existing annotation both in identifying actual intron-exon structure as well as in identifying different splice variants as was shown in maize where 16-17% of all transcripts were novel splice isoforms (Kakumanu et al., 2012). Furthermore, in contrast to the high background noise caused by cross-hybridization in micro-arrays most RNA-seq reads can be unambiguously mapped to a region of the reference genome. This makes RNA-seq an excellent tool to differentiate between isoforms of a gene family, which are a widespread phenomenon in complex crop genomes. On the other hand, the alignment of sequence reads that are shared between several loci and therefore align to several locations on the genome is still complicated. One solution is to assign these reads proportionally to the number of unique and splice reads at these loci (Mortazavi et al., 2008). Moreover, aside from being unbiased towards previous sequence knowledge, RNA-seq is also much more sensitive. This sensitivity comes at a price though. To detect rare transcripts, more coverage and therefore more sequencing depth is needed which increases the sequencing cost. Lastly, the dynamic range of RNA-seq is also substantially higher with about five orders of magnitude compared to the couple of hundred-folds of microarrays (Wang et al., 2009; Zhao et al., 2014).

1.3.2 The future for transcriptomics in crops

With the introduction of NGS, RNA-seq seems the transcriptomics tool for the future, especially in crops. At the moment, the high costs associated with RNA-seq still prevent the large scale analysis of many varieties in different conditions. However as the NGS techniques keep evolving, costs are likely to drop and might no longer be the limiting factor in the future. As more and more genomes are sequenced, alignments to reference genomes will probably become the norm, which also significantly reduces the analysis time required for *de novo* assembly. Nowadays, qPCR analysis is often combined with RNA-seq data to demonstrate the validity of the results obtained with RNA-seq in a selection of genes. qPCR can and will also play a complementary role to RNA-seq allowing the screening of the expression of candidate genes in more varieties and/or under different conditions.

1.4 Proteomics

Due to regulation at the translational level and post-translational modifications (PTMs), it is important to also analyze the final product of the genetic code, proteins, since they are the actual effectors in the plant. A discrepancy between mRNA levels and protein levels has been shown in many cases and mRNA contains no information on, for instance, PTMs which can affect a protein's function or activity (Gygi et al., 1999b). Whereas genomics and transcriptomics have the advantage of a homogenous sample substrate, the heterogeneity of proteins poses significant challenges for proteomic analysis. While DNA and RNA only contain 4 nucleotides, proteins are made up of 20 amino acids with very different chemical properties. Moreover, there is a significant difference in abundance between the most and the least abundant protein and since no amplification methods, such as polymerase chain reaction for DNA, exist for proteins, all methods show a reduced dynamic range. Current proteomics technologies can measure an order of magnitude of 5 in intensity, but usually only 3 orders of magnitude can be successfully identified (Michalski et al., 2011) (Figure 1.2).

It is therefore at the moment still impossible to use one single technique to study the whole proteome. We will first outline the high-throughput blind differential analysis techniques followed by the methods used to validate candidate genes and finally we will focus on a few subproteomes.

All proteomics approaches involve the same basic steps: protein extraction, separation and/or visualization of proteins and/or peptides, quantification and identification.

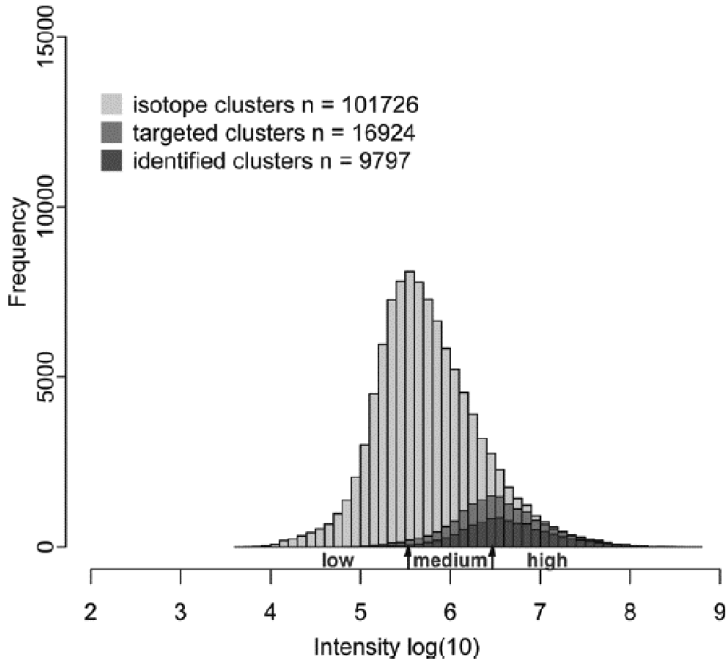

Figure 1.2: Histogram of intensity of the detected features in an LC-MS/MS run (figure from Michalski et al., 2011). An order of magnitude of 5 in intensity is measured at the MS level, but identified peptides only cover an order of magnitude of 3. Light gray: all peptides, mid gray: all peptides targeted for MS/MS and dark grey: targeted and identified peptides.

1.4.1 High-throughput blind differential analysis

To date the most used method in differential proteomics in crops is the analysis of whole cell extracts using a gel-based approach (Agrawal et al., 2013; Barkla et al., 2013). Gel-based proteomics has the reputation, however, of being a slow and cumbersome art. The development of the more high-throughput gel-free approaches in crops might provide an answer to some of the problems associated with gel-based proteomics.

1.4.1.1 Gel-based differential proteomics: 2DE

1.4.1.1.1 Protein extraction

The first crucial step in plant proteomics is the extraction of the proteins from the plant material. Plant cells have a low protein to volume content, but contain significant amounts of interfering substances such as phenols, polysaccharides, lipids, pigments, proteases and oxidative enzymes. The most used protocol is a TCA/acetone precipitation protocol which was developed in the eighties and has

more recently been adapted (Damerval et al., 1986; Rabilloud and Chevallet, 2000; Mechin et al., 2007). However for even more recalcitrant plant tissues, a phenol extraction followed by ammonium acetate precipitation was shown to be more powerful. This protocol was also developed in the eighties and more recently adapted (Schuster and Davies, 1983; Saravanan and Rose, 2004; Carpentier et al., 2005). If only a specific subgroup of proteins is targeted, e.g. membrane proteins, a pre-fractionation technique needs to be applied as is discussed in 1.4.3.

1.4.1.1.2 Protein separation

Extracted proteins are then separated by two-dimensional gel electrophoresis, first described by O'Farrell (1975), in which denatured proteins are first separated based on their isoelectric point (pI) and then in a second dimension based on their molecular weight.

1.4.1.1.3 Protein visualization and quantification

Several techniques exist to visualize the proteins on the gel. Post-electrophoresis stains include Coomassie brilliant blue, silver and the fluorescent SYPRO dyes. While all these post-electrophoresis stainings are limited to one sample per gel, multiplexing of two samples and an internal standard is possible using the difference gel electrophoresis (DIGE) technique (Alban et al., 2003). Samples are labelled prior to electrophoresis with dyes that have different excitation and emission wavelengths. While two samples are usually run per gel, more importantly a pooled internal standard is also added allowing for easier gel-to-gel matching and better quantification eliminating a significant amount of experimental variation. Fluorescent labels and specialized equipment such as low fluorescent glass plates, a fluorescence scanning system and specific software however do raise the costs. Quantitative analysis of the detected spot intensities is performed with specific image analysis software such as DeCyder (GE Healthcare Life Sciences[7]) and Delta2D (Decodon[8]). In brief, the software will detect spots on the images. Subsequently, a spot volume threshold will need to be determined by the user to separate the actual protein spots from erroneously detected noise signals. The software will then perform the matching of the gels based on the internal standard images followed by normalization procedures, quantification and statistical analysis of the standardized spot volumes. Fold changes in relative abundances are reported between the tested conditions. The statistical approaches include both univariate and multivariate methods such as Student's t-test and PCA.

[7] http://www.gelifesciences.com

[8] https://www.decodon.com

1.4.1.1.4 Protein identification

Until the end of the eighties, Edman degradation was the main method for protein identification. This method labels the N-terminal peptide, cleaves it from the rest of the protein and identifies it using chromatography (Edman and Begg, 1967). Several cycles of these steps will lead to the identification of the protein sequence. Nowadays, the identification of proteins in spots is mostly carried out using mass spectrometry. Selected spots are picked from the gel and the proteins in the gel plug are digested into peptides using a protease such as trypsin. The peptides are extracted from the gel plug and analyzed with a mass spectrometer. In general, determining only the m/z of the peptide fragments generates insufficient information for positive identifications in crops. Two mass analyzers separated by a collision cell are usually combined to form a tandem mass spectrometer (MS/MS). The first mass analyzer will determine the m/z of the peptides, the so-called precursor ions. Consequently, precursor ions with specific m/z ratios are selected and fragmented in the collision cell. Finally the intensity and mass of these fragments are measured in the second mass analyzer on the detector. The signal data are converted into mass spectra: one MS spectrum of the whole mixture containing all measured peptides (precursor ions) and one MS/MS spectrum per fragmented precursor ion. The latter will lead to peptide identifications and in combination with the former to protein identifications using search algorithms against amino acid sequence databases. The most popular search algorithms include Mascot (Perkins et al., 1999), SEQUEST (Eng et al., 1994), X!Tandem (Craig and Beavis, 2004) and OMMSA (Geer et al., 2004).

Matrix-assisted laser desorption/ionization (MALDI) followed by time-of-flight (TOF or TOF/TOF) mass spectrometry is still the most widespread identification method in plant proteomics following two-dimensional gel electrophoresis (2DE) analysis (Champagne and Boutry, 2013). The medium sensitivity and resolution of MALDI-TOF/TOF mass spectrometers are sufficient for most gel-based proteomics goals.

1.4.1.2 Gel-free differential proteomics: LC-MS/MS or 2DLC-MS/MS

The other main method in differential proteomics is the gel-free approach, also referred to as the bottom-up or peptide-based approach. After extraction, proteins are immediately digested into peptides and this peptide mixture is separated, using liquid chromatography (LC) or 2D-LC. The fractions are analyzed on-line by coupling the LC-column to a mass spectrometer. The digestion into peptides is recommended as high-throughput top-down mass spectrometry of intact proteins is extremely challenging.

1.4.1.2.1 Protein extraction

As in gel-based proteomics, a gel-free method starts with the extraction of the proteins. While the same extraction methods such as TCA and phenol extractions can be used, gel-free approaches need to be performed on detergent-free samples. Detergents can after all prevent enzymatic digestion and dominate mass spectra due to their easy ionization and relative abundance compared to peptides (Wisniewski et al., 2009). Therefore proteins must either be solubilized without detergents, an acid labile surfactant must be used as a detergent, or the detergent has to be removed by filter aided sample preparation (Yu et al., 2003; Manza et al., 2005; Wisniewski et al., 2009; Vertommen et al., 2011a). Protein samples are then digested into peptides using an enzyme such as trypsin.

1.4.1.2.2 Peptide separation

Separation of the peptides is performed by liquid chromatography. In the beginning of this century, the multidimensional protein identification technology (MudPIT) was introduced. The peptides are first separated based on their inherent charge on a strong cation exchange column (SCX). Fractions are then further separated based on the hydrophobicity of the peptides using a reversed phase (RP) column. The columns are coupled to a mass spectrometer which creates an automated, high-throughput workflow (Washburn et al., 2001). The MudPIT method has largely been replaced by long gradient RP or RP-RP coupled chromatography since mass spectrometers with higher mass accuracy, resolving power and scan speed are now available (see 1.4.1.2.3). In RP-RP, two different pH are used for elution from the first and the second column, respectively. Gilar and his colleagues demonstrated that a higher peak capacity in the first dimension compared to SCX resulted in more fractions with less overlap and also observed less peptide losses in that first dimension compared to SCX. Moreover, since both mobile phases are salt-free, they are compatible with MS analysis (Gilar et al., 2005). Our group was one of the pioneers exploring this technique on a crop (Vertommen et al., 2011a).

1.4.1.2.3 Peptide quantification and protein identification

Gel-free differential proteomics makes use of mass spectrometry for both quantification and identification.

Protein identification and quantification remains a challenge in gel-free approaches in plants. Peptides shared between several proteins do not contribute to the conclusive identification of a protein. This is the so-called protein inference problem. Tryptic specific peptides need to be measured and identified for final protein identification and quantification. Large gene families present in plants complicate the identification and quantification further.

The quantification of peptides/proteins is based on one of two approaches: label-based or label-free. In the label-based methods one can discern chemical labeling methods, which take place after extraction of the proteins, and metabolic labeling methods, which involve incorporation of the label during the growth of the plant. Isotope-coded affinity tags (ICAT) and isotope-coded protein labels (ICPL) are both quantified at the MS level (Gygi et al., 1999a; Schmidt et al., 2005). One sample is labelled with a tag with light isotopes and the other with a heavy isotope tag which results in predictable peptide-tag weights which are measured at the MS level. The isobaric tag for relative and absolute quantification (iTRAQ) method uses tags that generate isobaric precursor ions and produce specific fragments (reporter groups) at the MS/MS level (Wiese et al., 2007). Metabolic labeling methods on the other hand offer the advantage that all samples can be mixed even before protein extraction reducing the effect of technical variation considerably (Bindschedler and Cramer, 2011). Indeed, metabolic labeling already takes place within the organism during protein biosynthesis through the incorporation of labeled amino acids or important nutrients which are present in the growth medium. The most well-known method, stable isotope labeling with amino acids in cell culture (SILAC), has been shown to be less successful in plant systems than in animal systems as labeled amino acids were only partially incorporated in cell cultures (Ong et al., 2002; Gruhler et al., 2005). Plants as autotrophic organisms are able to synthesize all amino acids from inorganic nitrogen in the medium and therefore do not show full incorporation of what is an essential amino acid in animals. On the other hand, hydroponic isotope labeling of entire plants (HILEP) and stable isotope labeling in planta (SILIP) use the plant's incorporation of inorganic ^{15}N into amino acids to label the proteins and accomplish a more complete incorporation (Bindschedler et al., 2008; Schaff et al., 2008). All inorganic nitrogen sources in the form of salts in the media (HILEP) or in the fertilizer for the soil (SILIP) are the standard ^{14}N sources in one sample and are replaced with ^{15}N-labelled nitrogen sources in the other sample. HILEP was developed in *Arabidopsis* but the same authors show that proteins in woodland strawberry can also be labeled this way (Bindschedler et al., 2008; Agrawal et al., 2013). The SILIP method was shown to work with two month-old tomato plants grown in greenhouses (Schaff et al., 2008). Label-free LC-MS/MS approaches on the other hand can be performed on any biological sample and experimental set-up, require less time-consuming steps and avoid the cost of the labeling reagents (Bantscheff et al., 2007). The two widely used quantification methods in label-free quantification are based on either measuring the intensities of the precursor ions or spectral counting. Spectral counting is a very simple procedure and is based on the idea that more MS/MS spectra of peptides of a more abundant proteins will be collected than of less abundant proteins (Washburn et al., 2001). A more advanced method called protein abundance index (PAI) calculates the number of sequenced peptides

belonging to a protein divided by the theoretical number of peptides and takes into account that a larger protein and proteins with more peptides within the measured mass range generate more observed peptides (Rappsilber et al., 2002). The same authors later refined the method by introducing the exponentially modified PAI (emPAI) which is calculated as 10^{PAI}-1 and which was shown to be directly proportional to the protein content (Ishihama et al., 2005). The alternative label-free method to spectral counting, which uses peak intensities, extracts the ion chromatogram of each peptide from the LC-MS/MS run and integrates the peak areas over the time the peptide was eluted (Chelius and Bondarenko, 2002). For label-free quantification methods the main bottleneck remains the variability introduced during the chromatography step. As only one sample can be run each time, different chromatography runs need to be aligned to each other to find corresponding peptides. A highly reproducible peptide chromatography profile is therefore necessary.

Most mass spectrometry approaches use data-dependent acquisition (DDA) analysis in which the precursor ions are selected and fragmented based on the data from the MS scan. DDA has a strong bias towards more abundant peptides since precursor selection is based on the ion intensity and charge determination determined during the MS survey scans. In most MS, the precursor survey scan and the MS/MS fragmentation scans duty cycles are performed in series (Bantscheff et al., 2007). A data-independent acquisition approach (DIA), called MS^E, was developed for label-free approaches by using a mass spectrometer which analyzes all peptides at once in a certain chromatographic window by applying an alternating energy level to the collision cell. At low energy levels, the precursor masses are measured and at high energy levels all precursor masses are fragmented at once (Plumb et al., 2006). This enables a very fast cycling between low and high energy duties, enabling an accurate quantification of the precursor ions. The fragmentation spectra however are much more complex and this leads to poor protein identifications. This is further exacerbated by the presence of large gene families within the genome as well as well as the heterozygous and polyploid nature of many crops. A combination of DDA-based spectral libraries and quantitative DIA approaches has successfully been used by our group to characterize changes in the proteome during the storage of the apple fruit (Buts et al., 2014). Ion mobility is now also used in combination with MS^E, resulting in high definition MS^E. Ion mobility uses the differences in structural properties of precursor ions to give those precursor ions a different kinetic energy before fragmentation. The different mobility of the precursor ions allows to calculate the origin of the fragment ions and will lead to less chimeric spectra and more confident peptide identifications (Valentine et al., 2001). This approach however

suffers from transmission loss and detector saturation and consequently hinders the quantification of both high and low intensity peptides (Shliaha et al., 2013).

1.4.1.3 Gel-based vs gel-free differential proteomics

Both gel-free and gel-based methods have their advantages and disadvantages. 2DE has the advantage of being a high resolution technique which is able to resolve protein isoforms and proteins with PTMs as is discussed in 1.4.2.1. As more and more crops are being sequenced, gel-free, peptide-based proteomics will likely become the standard in differential studies as it offers a more high-throughput analysis. As stated before neither technique can quantify and identify the whole proteome in a cell. In a gel-based approach only 1000-3000 spots will be visualized on a 24 cm gel. According to Wilkins et al. (1998), only the most abundant proteins, which are present in more than 10,000 copies in a cell, are visualized. Hydrophobic and basic proteins are also difficult to detect on gels (Wilkins et al., 1998). Moreover, gel-based proteomics is also restricted in the size of proteins that can be analyzed. High molecular weight proteins are badly transferred from the strips used for the first dimension separation to the gels used for the second dimension separation. They are also not well resolved on the second dimension of 2DE gels although the use of gradient gels can improve this resolution. Low molecular weight proteins (<10 kDa) on the other hand will co-migrate with the SDS front and cannot be resolved either in the classical buffer system. Gel-free or peptide based approaches also struggle with low abundant proteins. As stated above, the major disadvantage of the peptide-based approach lies in the disconnection between a protein and its peptides. A protein sample containing several thousands of proteins is digested and all these peptides are now analyzed at once. This leads to both identification and quantification problems in the case of non-sequenced organisms such as most crops as discussed before in 1.4.1.2.3. As genomes become available, however, this identification problem can be tackled more efficiently.

1.4.2 Digging deeper into differential proteins

1.4.2.1 Analyzing protein species

While 2DE may no longer be the tool of choice in high-throughput differential proteomics in the future, it is still very effective to identify and quantify protein species caused by genetic variations, alternative splicing and/or PTMs. Several spots on 2D gels often get the same general identification but in a crop with a sequenced genome it becomes feasible to identify isoforms, alternative splice variants and/or PTMs (Chapter 5). Protein species, when caused by sequence variation or PTMs, often have small differences in pI and a similar mass (Carpentier et al., 2011; Henry et al., 2011). This small difference in pI can be observed on 2D gels, especially on

zoom strips which offer a high resolution. On the other hand a modification like glycosylation might result in the same pI, but causes a mass shift that is distinguishable on gels (Laugesen et al., 2007). The resulting horizontal or vertical trails of spots are easy to detect on a 2D gel and the quantification of these spots is very straightforward which makes differentially expressed isoforms or important PTMs simpler to observe. However, not all modifications or amino acid replacements result in differences in the net charge of proteins or mass and therefore several isoforms can be present in the same spot (Chapter 5). Moreover, 2D does not give any information on the position of the PTM on the protein. Therefore, a single spot can for example contain two forms of a modified protein with the same type of modification on different sites of the protein (Rogowska-Wrzesinska et al., 2013). Because of the direct visualization on the gels, however, it is possible to evaluate the minimum number of isoforms present. The complexity in a single spot is also much lower than in completely peptide-based proteomics using a gel-free blind high-throughput approach where finding the low-abundant different or modified peptide is similar to looking for a needle in a haystack. However, to really dig into the make-up of the spots at the isoform and/or PTM level, more advanced mass spectrometry technology than MALDI-TOF/TOF MS, the usual mass spectrometer linked to 2DE, might be necessary (Chapter 5). Further separation of the peptides in a spot during liquid chromatography followed by MS/MS allows for more peptides to be identified and will lead to a bigger coverage of the proteins. This strategy is obviously no longer high-throughput and once the exact peptide which identifies the isoform or PTM has been found, a more high-throughput method to analyze and verify quantitative expression over time or in more varieties can be used, as is discussed in the next paragraph.

1.4.2.2 Validating differential proteins

Specific peptides, and therefore the proteins to which they belong, can be very accurately monitored using selected reaction monitoring (SRM), also named multiple reaction monitoring (MRM). The specificity of this technique allows the identification of low abundant peptides and proteins in complex mixtures (Lange et al., 2008). A mass spectrometer is set to selectively detect one or several specific precursor ions and a selection of fragment ions. The combination of a peptide and fragment for a specific protein is called a transition. Each transition needs to be carefully developed so that the selected peptides have a good MS response and are specific for a certain isoform or modification so proteotypic peptides need to be selected. Moreover for each proteotypic peptide fragment ions need to be determined that provide a good signal intensity as well as discriminate the targeted peptide from the other peptides (Lange et al., 2008). The sample is often spiked with a known amount of heavy-labeled target peptide to provide absolute quantification (Wienkoop et al., 2010).

The technique has also been dubbed Mass Western as it is able to outcompete immunoassays in the study of isoforms once efficient proteotypic peptides have been identified (Lehmann et al., 2008). Four different sucrose phosphate synthase isoforms were quantified in *Arabidopsis* using SRM (Lehmann et al., 2008). The different isoforms showed differential expression patterns in different tissues and the authors also identified a cold responsive isoform. SRM can be used to very accurately identify specific isoforms in a multitude of conditions, tissues and experiments. The scope of a SRM study can stretch from a few specific marker proteins to complete metabolic pathways (Wienkoop et al., 2010). While the standard peptide-based approaches are high-throughput as to the number of proteins in an experiment, only a limited number of samples can be run due to long run times and inter-run variation. SRM/MRM approaches on the other hand focus on fewer proteins but the shorter run time allows for more sample to be run providing a higher throughput in number of biological replicates and/or conditions. Consequently, the use of 2DE combined with LC-MS/MS to analyze protein isoforms in detail combined with SRM once the relevant isoforms and proteotypic peptides have been identified, provides an excellent approach for both detailed identification and validation of important protein candidates discovered in blind high-throughput analyses.

1.4.3 Subproteomes

Several protein groups, such as low abundant proteins, membrane proteins and modified proteins are severely underidentified in the proteomic analysis of total cell extracts. Traditional prefractionation and enrichment methods might introduce additional variability in the samples and alternative methods are being introduced (Barkla et al., 2013; Vanderschuren et al., 2013).

Low abundant proteins such as regulatory proteins which may play key roles in the response to abiotic stress are often difficult to detect due to the presence of other more abundant proteins. Especially in green plant tissue, ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) dominates the protein content (Ellis, 1979), visually obscuring several spots in gel-based approaches. In gel-free approaches, peptides co-eluting with RuBisCO peptides could be less ionized due to ion suppression and are less likely to be picked for MS/MS analysis due to the overabundance of RuBisCO peptides. Combinatorial peptide ligand library is a method which removes the most abundant proteins before protein/peptide separation (Thulasiraman et al., 2005). Beads carry a large number of copies of the same ligand, often a hexapeptide. Each bead carries a different hexapeptide which results in 64 million combinations when all 20 natural amino acids are used (Righetti et al., 2006). While highly abundant proteins will cover their bead quite quickly and

the protein surplus can no longer bind and is removed, lower abundant peptides will continue binding to their bead when more of the sample is added. This will selectively enrich for the lower abundant proteins. While this technique has mostly been used on human tissue and biological fluids, it has successfully been applied to *Arabidopsis* and spinach leaves as well as pumpkin phloem (Boschetti et al., 2009; Fasoli et al., 2011; Frohlich et al., 2012). While the enrichment may vary from protein to protein within a sample, it was shown that a given protein is proportionally retained in different samples treated in parallel in human plasma and red blood cells (Roux-Dalvai et al., 2008; Sihlbom et al., 2008). This differential comparison has not yet been performed in plants.

Another group of particular importance are membrane proteins. Membrane proteins are poorly soluble and are also low abundant (Vertommen et al., 2011b). Centrifugation-based separation consists of several centrifugation steps leading to a membrane enriched fraction which is often followed by density gradient centrifugation to isolate specific membranes or organelles. Another method is free-flow electrophoresis in zone electrophoresis mode which separates membranes based on their charge (Braun et al., 2007). Combining these two approaches results in a substantial gain in purity of membrane fractions but specialized equipment is required (Eubel et al., 2005; Vertommen et al., 2011b). A peptide-based approach is favored as the solubility problem of membrane proteins can be circumvented by focusing on the soluble peptides (Vertommen et al., 2011b).

Modified proteins are another important target group in proteomics. Identifying and mapping PTMs, determining their location on the protein, in crops is still in its infancy. The most well-known and extensively studied modification is phosphorylation. A recent review by Rampitsch and Bykova (2012) showed that most phosphoproteomics abiotic stress research has focused on *A. thaliana* and rice and to a lesser extent on maize and soybean. Gel-based methods usually rely on phosphospecific stains or immunodetection while in gel-free methods the enrichment of phosphopeptides is usually achieved through immunoprecipitation, immobilized metal-ion affinity chromatography or titanium dioxide chromatography. Mapping a PTM requires extensive sequence information but as more and more crops species are being sequenced, the necessary genomic information for the identification of phosphopeptides and other PTMs will become available (Rampitsch and Bykova, 2012). The transient nature of many PTMs however remains a challenge. Bond et al. (2011) and Remmerie et al. (2011) gave an extensive overview of the available methodologies for the study of amongst others phosphorylation, glycosylation and ubiquitination. Identifying and mapping PTMs on peptides and proteins is the first step in PTM research. Next, SRM, as described

above, can be used on known modified peptides to quantitatively measure the abundance among a range of conditions and varieties.

1.4.4 The future for proteomics in crops

Proteomics will always struggle to quantify all proteins in a sample. Both the protein chemical heterogeneity and the broad dynamic range make it impossible to use one technique to analyze it all. Rather, the best proteomics approach is determined by the current biological question. The ideal proteomics approach would be a combination of high-throughput systems that offer the ability to study the differential expression of as many proteins as possible followed by validation in more varieties and conditions. But specific workflows need to be set up to study e.g. membrane proteins, low abundant proteins, isoforms and PTMs (Chapter 5).

1.5 Metabolomics

Metabolomics is defined as the comprehensive analysis of the metabolites of a biological system (Fiehn, 2002). No single analysis method can identify and quantify all of the metabolites present in a cell and is unlikely to be ever developed. The plant kingdom is estimated to contain anywhere between 100,000 and 1,000,000 metabolites (Sardans et al., 2011; Obata and Fernie, 2012). But numbers up to 5 million structures have been named (Weckwerth, 2011). For a single species the number is estimated at a few thousand metabolites with an estimate of ca. 5000 metabolites for *Arabidopsis* (Obata and Fernie, 2012). Metabolites have a greater chemical diversity compared to that of nucleic acids and proteins (Fiehn, 2002). Plant metabolite profiling approaches are often used and focus on a specific subset of metabolites.

1.5.1 Metabolomics technologies

Gas chromatography-mass spectrometry (GC-MS) remains one of the most used methods in metabolomics approaches (Obata and Fernie, 2012). It can be used to analyze volatile components or components that can be volatilized such as sugars, sugar alcohols, amino acids, organic acids and polyamines, and therefore focuses on the primary metabolism pathways. The short running time and low cost as well as the existence of standard protocols and several metabolite databases for peak annotation are the main advantages of this technique. However the use is limited to thermally stable volatile metabolites excluding thermolabile and large molecules and usually a derivatization step is required before analysis. LC-MS using a reverse phase column is frequently used to study secondary metabolites because it allows the

study of many metabolites at once (Obata and Fernie, 2012). The introduction of ultra-performance liquid chromatography resulted in a higher resolution, sensitivity and throughput compared to the standard high-performance liquid chromatography (Sardans et al., 2011; Obata and Fernie, 2012). The liquid chromatography columns are coupled to high performance mass spectrometers for identification of the compounds. Fourier transform based mass spectrometry delivers the highest accuracy but other mass spectrometers are frequently used as well (Sardans et al., 2011; Obata and Fernie, 2012). The LC-MS approach allows the study of many high molecular mass and thermolabile compounds at once. There are, however, more difficulties in establishing spectral databases as protocols are less standardized and the retention time in the LC phase is different for each instrument. Both gas chromatography and liquid chromatography methods are mostly limited to targeted or profiling analysis and to obtain a true metabolomics approach unknown compounds should be able to be identified too. Nuclear magnetic resonance (NMR) has proven to be an appropriate tool for determining the structure of novel compounds. This approach is highly accurate and reproducible but less sensitive than the previous methods so only abundant metabolites will be measured (Fiehn, 2002; Sardans et al., 2011; Obata and Fernie, 2012). At the moment even the combination of all these methods, similarly to the situation in proteomics, cannot cover the whole range of metabolites.

1.5.2 The future for metabolomics in crops

While in many of the other omics approaches crop studies are still implementing analysis techniques already available to model plants, metabolomics is probably the only approach which is readily applied to crops. On the other hand, due to the immense heterogeneity in metabolites, it is still very difficult to get a complete overview of all metabolites and the abundance problem is not solved either.

1.6 Phenomics

Plant phenotyping is a crucial step for stress research. The speaking plant concept was first introduced by Hashimoto et al. in the 1980s in which he stated one could set up a system in which the input were the environmental factors in a greenhouse and the output the plant responses or speaking plant as he called it (Hashimoto et al., 1984). To this day we are still analyzing what plants ‘say’ when subjected to stress. Commonly used phenotyping tools harvest (parts of) the plant to measure weight and leaf area. Samples are then collected for e.g. determination of dry matter to estimate plant water content and carbon isotope determination to estimate lifetime stomatal closure (Furbank and Tester, 2011). These methods need

destructive sampling at selected times and are usually labor-intensive and therefore fewer replicates are observed (Walter et al., 2007; Furbank and Tester, 2011). This often leads to the decision to only analyze the growth or yield at the end of an experiment (Furbank and Tester, 2011). To relieve the phenotyping bottleneck, more high-throughput environment controlled platforms need to be developed to improve precision and to eventually reduce the need for replications in the field (Walter et al., 2007; Furbank and Tester, 2011; Dhondt et al., 2013).

1.6.1 Phenomics technologies

In the last years, a number of high-throughput image-based technologies have been developed offering the researcher non-destructive methods to observe a plant's response over time. In 1999, Leister et al. were the first to describe a non-invasive image analysis to perform a large scale evaluation of plant growth in *Arabidopsis thaliana* (Leister et al., 1999). They showed that 'plant area estimation' correlated well with plant fresh weight for certain ecotypes. The first fully automated platform for *Arabidopsis* to perform image analysis and apply drought stress by weighing and adding water automatically was established in 2005 under the name PHENOPSIS (Granier et al., 2006). A year later an automatic image acquisition system GROWSCREEN showed that also *Nicotiana tabacum* growth could be assessed using projected leaf area (Walter et al., 2007). Over the last years several commercial platforms using conveyor belts have been developed for controlled environments and are in use all over the world. Examples include the ScanAnalyzer platform from Lemnatec[9] and the PlantScreen™ from Photon System Instruments[10]. These platforms integrate weighing and watering systems with several imaging modules such as visible light, near-infrared, infrared and others which can be used on both roots and shoots. These phenotyping systems are often modular and can be adapted for the type of research being performed, going from osmotic, drought and salinity stress to nutrient stress and biotic stress screens.

Visible light imaging mainly measures parameters such as leaf area, leaf orientation, plant height and width to estimate amongst other biomass development and plant architecture, but additionally can be used to assay e.g. leaf color and senescence. Infrared imaging identifies the leaf temperature which is correlated to the opening and closing of stomata. Additionally fluorescence imaging can be used to measure the chlorophyll fluorescence which is a parameter for the efficiency of photosynthesis (Jansen et al., 2009; Furbank and Tester, 2011). In the future several

[9] www.lemnatec.de

[10] www.psi.cz

other spectral measurements will be added such as near-infrared imaging to estimate water levels in leaves as well as in the pots to investigate moisture distribution in the root column.

In the field, hyperspectral remote sensing is used as it is a fast tool which offers good spatial and temporal resolution with main applications in prediction of chlorophyll content, leaf area index estimations and water status detection (Stuckens et al., 2011).

Root phenotyping remains a bottleneck in the automated approach. Root systems are often manually laid out in front of a camera or on a scanner but these are often destructive approaches (Dhondt et al., 2013). Nagel and colleagues have developed two non-destructive systems one using petri dishes filled with an agar medium and the other a rhizotron system with soil (Nagel et al., 2009; Nagel et al., 2012). Using specifically developed software programs, GROWSCREEN-Root and GROWSCREEN-Rhizo, they greatly advanced the automation of non-invasive root architecture software, tracking root length and branching rates and angles.

One of the risks of high-throughput and fully automated workflows from image acquisition to image processing, is data quality deterioration. Several checkpoints should be in place and although supervised image processing has a negative connotation because the workflow is no longer fully automated, it does provide the chance to review and correct errors in obtained data. For processing steps that cannot be fully performed yet by a computer, a semi-automated workflow is the best option (Dhondt et al., 2013).

While high-throughput platforms have been established in several universities, research centers and even commercial biotech companies, they are still very costly and their application to crops remains rather limited. Aside from *Arabidopsis* and small rosette plants, the automated systems used in controlled environments are mainly applied to cereals such as rice and barley and are less suitable for other crops as some plants cannot be moved or are too large. This does not prevent several advancements for other crops in image-based analysis. Tall greenhouse plants such as pepper are for example difficult to transport to a dedicated imaging chamber. Van der Heijden and colleagues therefore constructed a device equipped with several cameras to span the 3 m high plants which automatically take pictures of the plants (van der Heijden et al., 2012). A creative solution to the high cost of 3D measurements of plants was the use of the depth camera from the Microsoft© Kinect system on rosebushes, yucca plants and small apple trees. This system based on reflected light rather than time-of-flight like most depth cameras has a lower spatial and depth resolution than these classical cameras but is able to monitor

several phenotypic traits such as leaf curvature, leaf morphology and orientation (Chene et al., 2012).

1.6.2 The future for phenomics in crops

We can conclude that automated, high-throughput phenotyping is still mainly limited to cereals or small rosette plants. However, research groups dedicated to a specific crop start to apply image-based techniques and work on specific solutions to automate their work on that crop.

1.7 Integrations of omics approaches

1.7.1 Linking of transcriptome and proteome to genome

The link between transcriptome or proteome on the one hand and genome on the other hand is often a one-way street. The genome is used to annotate results obtained by transcriptomics or proteomics. Especially in proteomics, cross-species identification in crops is still quite common, but this will be replaced as more genomes become sequenced. However proteome and transcriptome data are increasingly used to annotate the genome as well. As stated above, RNA-seq provides the information necessary to identify previously unknown gene coding sequences as well as to improve existing annotations both in identifying actual intron-exon structures and the different splice variants. The same can be done with proteomics data. Renuse et al. (2011) define 'proteogenomics' as the correlation of proteomic data with genomic and/or transcriptomic data to enhance our understanding of the genome. MS/MS data are searched against a six-frame translation of the genome or transcriptome with the goal of identifying novel peptides and consequently proteins. Alternatively, *de novo* sequencing results, in which the amino acid sequence of a peptide is determined based on the spectra, can be searched against the genome or transcriptome to find these novel peptides. Proteogenomics efforts have already delivered results in *Arabidopsis*, rice and maize. In *Arabidopsis* results suggested that 13% of the proteome was not previously covered due to missing or incorrect gene models. Using about a third of the novel identified peptides, 778 new protein-coding genes were identified and 695 gene models were updated (Castellana et al., 2008). A similar strategy was used in *Zea mays* and resulted in the identification of 165 new protein-coding genes and the refinement of 741 gene models (Castellana et al., 2014). A proteome database for rice, OryzaPG-DB has been released. At release, it contained 3200 genes of which 40 with new gene models. All these gene models are based on experimental shotgun

proteomic data (Helmy et al., 2011). One of the main bottlenecks in the use of proteogenomics is the data analysis. While the search against a database of known or predicted proteins is relatively straightforward, large databases are generated for the six-frame translated genome resulting in higher error rates. High quality data are therefore very important in proteogenomics. The development of a dedicated software for an entire proteogenomics analysis would increase the genome annotation of several organisms (Renuse et al., 2011).

1.7.2 Systems biology

Systems biology is defined as the integration of experimental data from different omics platforms, the genome scale reconstruction of entire metabolic pathways and the derivation of models capable of predicting the phenotype of plants in their environment (Weckwerth, 2011). Not surprisingly systems-based approaches are most advanced in *Arabidopsis*. Several studies already integrate multiple omics approaches into interaction networks (Wienkoop et al., 2008; Sulpice et al., 2010; Araújo et al., 2012; Baerenfaller et al., 2012; Higashi and Saito, 2013), but the integration of all omics approaches (genomics, transcriptomics, proteomics, metabolomics and phenomics) into one study has to our knowledge not yet been performed. Major challenges remain to successfully implement a complete systems biology approach. As of yet, bio-informatics tools still need to be developed to integrate and statistically analyze data generated by all the omics approaches. The biggest challenge however is to present these data in a way that is comprehensible (Mittler and Shulaev, 2013). Moreover, different plant parts, tissues and even cells probably respond differently and this will all need to be accounted for in the models (Mittler and Shulaev, 2013). This makes systems biology an incredibly powerful but expensive and complex approach.

Chapter 2

Abiotic stress research in crops using omics approaches, osmotic stress and banana in the spotlight

2.1 Introduction

The main research area of this dissertation is osmotic stress in banana. In the first part of this chapter, drought responses and osmotic stress in plants are discussed. In the second part, recent applications of the omics approaches, which were reviewed in Chapter 1, to abiotic stress research in crops are discussed. A more in-depth look into the omics approaches used in banana research, in general and specifically in abiotic stress studies, is taken in the third part.

2.2 Drought and osmotic stress

2.2.1 Drought stress and water deficit

Both in nature and in agriculture, plants, being sessile, often experience some form of stress imposed by external factors which negatively influence their growth, development and production (Bray et al., 2000). Biotic stresses are caused by another organism such as bacteria, fungi, nematodes and insects. Abiotic stresses, on the other hand, are caused by non-living environmental factors with drought, cold, heat, oxygen deficit due to waterlogging and nutrient deficiency as the most well-known examples (Bray et al., 2000).

The terms drought stress and water deficit are often used interchangeably by plant physiologists. Some authors, however, define drought as a meteorological term which usually points to a prolonged period of abnormally low rainfall (Passioura, 2007). Drought causes several stresses including temperature stress, light stress and water deficit with this last one being the most characteristic (Verslues et al., 2006). The term plant water deficit is therefore sometimes preferred over drought stress in research as often only this stress factor is actually studied. Plant water deficit is defined as the condition when a plant's water demand exceeds water availability (Pardo, 2010). Plant water deficit therefore is not only a component of drought stress but also of salinity and cold stress (Verslues et al., 2006). While water deficit also plays a role in these last two stresses, it is most important in drought.

This PhD research has however not focused on the application of actual drought stress by withholding water or limiting the water supplied to the plant. Rather an osmotic stress treatment was used to create a water deficit. This has the advantage that the level of the stress can be tightly controlled in a precise and reproducible manner (Verslues et al., 2006). Moreover, banana *in vitro* heterotrophic and autotrophic plant systems are much smaller and can be kept in growth chambers with temperature, humidity and light controls which provide a more stable

environment than greenhouses and fields. On the other hand, an artificial environment is created in which roots are not actually in a drier environment but rather are standing in a solution that is restricting the water availability. Although the responses to osmotic stress which causes a water deficit are similar to the water deficit caused by drought stress, experiments in true drought conditions with more realistic soil systems are needed to validate osmotic stress research (Skirycz and Inzé, 2010).

2.2.2 Drought tolerance mechanisms

Drought 'tolerance mechanisms' are usually divided into three response mechanisms: escape, avoidance and tolerance.

Drought escape is practiced by plants who complete their whole life-cycle before the onset of drought (Ingram and Bartels, 1996). Avoidance mechanisms include extensive root systems, reduced stomatal conductance and reduced leaf area. All these mechanisms are intended to avoid water deficit in the plant tissues by more absorption of water or minimization of water loss and are usually achieved through morphological changes (Reddy et al., 2004). Tolerance mechanisms on the other hand are aimed at keeping the plant functional during water deficit in the plant. This includes osmotic adjustment and the expression of several genes aimed at keeping the cell functional as is further discussed in 2.2.3. The most extreme example of drought tolerance can be found in the so-called resurrection plants such as *Selaginella lepidophylla* and *Craterostigma plantagineum* (Ingram and Bartels, 1996).

In practice, plants and crops do not use a single response mechanism but rather a combination of the previously described mechanisms.

Most of these tolerance mechanisms come at a price though. Short life cycles often lead to smaller yields in crops. Drought avoidance mechanisms often have a reduced carbon dioxide assimilation as they try to avoid water loss. Osmotic adjustment requires energy to produce the solutes and these have potentially toxic effects when acquired in high concentrations (Mitra, 2001).

2.2.3 Response to water deficit

Plants respond to water deficit using an array of physiological, biochemical and molecular responses. We highlight the ones most relevant to the performed research.

The first step in the plant response is the perception of the stress. A physical stress, e.g. lack of water, needs to be converted into a biochemical response, the triggering of a cellular transduction pathway. While it has long been suggested that plants have a two-component osmosensor similar to yeast to detect water deficit (Bray, 1997), the first pathways in water deficit sensing remain unclear (Miyakawa et al., 2013). More is known about the downstream signaling in stress. The plant hormone abscisic acid (ABA) is one of the most important signaling molecules during water deficit. In response to water deficit, ABA synthesis can be increased, its breakdown decreased, and stored ABA can be released resulting in a cytoplasmatic detection and a triggering of an ABA-mediated response. In guard cells this will result in stomatal closure to prevent water loss and in other cells this will trigger ABA-mediated transcriptional regulation (Shinozaki and Yamaguchi-Shinozaki, 2000). Both ABA-dependent and ABA-independent pathways are involved in the transcriptional regulation to water deficit. ABA-inducible genes contain a *cis*-regulating ABA-responsive element (ABRE) with an ACGT core (Bray, 1997). The most well-known *cis*-acting element found in ABA-independent pathways is the drought responsive element (DRE) which is activated by the transcription factors DRE-binding proteins (DREBs). Roychoudhury and colleagues as well as Golldack and colleagues have written in-depth reviews of the signaling pathways involved in water deficit (Roychoudhury et al., 2013; Golldack et al., 2014).

The closing of stomata is one of the major responses in drought stress to limit water loss through transpiration, but also directly limits the carbon dioxide uptake. This results in a diminished photosynthesis rate and growth reduction in the plants. Excess light energy which is no longer efficiently utilized in photosynthesis will generate reactive oxygen species (ROS) which cause damage to DNA and RNA, oxidize proteins and damage membranes (Ingram and Bartels, 1996). Whereas photosynthesis rates systematically decline during stress, the respiration rate shows more diverse stress-, organ- and species-specific responses (Flexas et al., 2005). Several studies have described decreased respiration rates in leaves, shoots and roots whereas others have shown unaffected or even increased respiration rates in water-stressed plants. These contradictions have not yet been resolved (Flexas et al., 2005). A higher respiration in the mitochondria is probably related to maintaining ATP synthesis to compensate for reduced photosynthesis. This will however also increase ROS production in mitochondria during drought (Miller et al., 2010). ROS have been shown to play a signaling role in the plant for stress (Golldack et al., 2014). To counteract the deleterious consequences of ROS, the expression of genes involved in ROS scavenging is induced. These proteins will either enzymatically reduce the ROS (e.g. superoxide dismutase and catalase) or are involved in the

production of antioxidantia which will reduce the ROS (e.g. glutathione reductase and glutathione transferase) (Mittler, 2002).

Osmotic adjustment is often reported as an important reaction to water deficit. This includes accumulation of sugars, amino acids, polyamines and quaternary amines (Reddy et al., 2004). These solutes lower the water potential in a cell which allows for longer absorption of water and therefore increased turgor pressure in the cell. These solutes probably have other roles as well including stabilization of proteins and membranes and ROS scavenging (Bartels and Sunkar, 2005). Osmotic adjustment has however been questioned for use in field crops as a positive correlation has only been shown under severe water deficits and overexpression has been shown to have pleiotropic effects and a stress-inducible and/or tissue-specific overexpression will be necessary (Serraj and Sinclair, 2002; Wang et al., 2003).

Aside from the above described mechanisms which often activate pathways in stressed plants proteins involved in cellular protection are often more abundant during stress. These include late embryogenesis-abundant proteins, aquaporins and chaperones (e.g. heat shock proteins) (Ingram and Bartels, 1996; Bartels and Sunkar, 2005).

2.2.4 Water deficit research

Traditionally three main approaches have been used in drought stress and water deficit research (Ingram and Bartels, 1996). The first one involves the analysis of desiccation tolerant systems such as seeds and resurrection plants. While surviving extreme desiccation is certainly a valuable strategy in nature, the relevance of researching severe stress has been questioned in the search for drought tolerant crops. The second strategy involves the use of a model plant with a large amount of genetic resources. In this PhD research, we focused on the third strategy by using the actual crop and analyzing its response to stress.

2.3 Omics for abiotic stress: recent applications

2.3.1 Genomics

The sequencing of crop genomes undoubtedly contributes in a major way to the advancement of abiotic stress research on crops. A reference genome facilitates both proteomic and transcriptomic research and the sequencing of additional varieties opens perspectives for GWAS. The number of GWAS on abiotic stress is still rather limited but GWAS on barley and rice have already been performed (Huang et

al., 2010; Cai et al., 2013; Visioni et al., 2013). The phenotyping and whole genome resequencing of 517 rice varieties for instance resulted in the identification of four loci which contribute to the phenotypic variance in drought tolerance (Huang et al., 2010).

2.3.2 Transcriptomics

The application of RNA-seq to study crop responses to abiotic stress has really taken off in 2013. Analyzed crops include potato, tomato, rice, Jatropha and soybean which had expected results including induction of transcription factors and stress response transcripts to stresses such as drought, exposure to exogenous ABA, cold and phosphorus deficiency (Vidal et al., 2012; Oono et al., 2013a; Wang et al., 2013a; Wang et al., 2013b; Zhang et al., 2014). The somewhat lesser known crops agave and horse gram, a potential supplier for bio-energy which is drought tolerant and a legume mostly used as fodder, do not have sequenced genomes yet and *de novo* transcriptome assembly proved to be successful resulting in 35,000 and 22,000 estimated protein coding genes, respectively (Bhardwaj et al., 2013; Gross et al., 2013). Since the genome of wheat is also not yet fully completed, a *de novo* transcriptome assembly was also used for wheat to study phosphorus deficiency (Oono et al., 2013b). Comparisons to other species such as rice and *Arabidopsis* showed that many features were well conserved and that RNA-seq was able to capture the phosphorus deficiency transcriptome profile of wheat. In alfalfa (*Medicago sativa*), a salinity stress transcriptome revealed that although it has a close relative in *Medicago truncatula*, only 50% of the differentially expressed genes identified in the alfalfa RNA-seq study are represented on an *M. truncatula* microarray and can correctly bind to the microarray probes (Postnikova et al., 2013). But RNA-seq can certainly be useful to discover new transcripts in species with a sequenced genome as well. Kakumanu et al. (2012) revealed that out of all transcripts 16-17% were novel splice isoforms in a maize drought transcriptome analysis. In the comparison of both reproductive and leaf meristem tissues, they found that in the drought-stressed maize ovaries programmed cell death was activated, the cell cycle was halted and drought signaling was impaired whereas carbon starvation signaling was heightened. These events were not observed in the leaf meristem where the antioxidant defense mechanisms seemed to function successfully and changes related to programmed cell death did not occur. While these RNA-seq results are very interesting, several of the identified transcripts belong to genes of unknown function. This is a widespread phenomenon and the same orthologs of unknown function are frequently found in drought related studies as shown by Dugas et al. (2011). The fact that these unknown gene products are so well conserved suggests that they may play an important role in the response to

water-limiting environments. In conclusion, RNA-seq offers researchers the ability to investigate species-specific abiotic stress responses in both unsequenced and sequenced crops, but the actual function of some of the identified transcripts remains elusive although a first annotation might be deduced from the experimental set-up.

2.3.3 Proteomics

As evidenced by the large number of reviews, abiotic stress responses in crops are often studied with proteomics approaches (Jorrín et al., 2007; Salekdeh and Komatsu, 2007; Jorrin-Novo et al., 2009; Roy et al., 2011; Sobhanian et al., 2011; Abreu et al., 2013; Agrawal et al., 2013; Barkla et al., 2013; Hossain et al., 2013; Ghosh and Xu, 2014; Ngara and Ndimba, 2014). Many of the recent studies have still employed 2DE approaches and several stress-responsive proteins are commonly identified in several treatments and species (Barkla et al., 2013). Proteins involved in energy metabolism and oxidative stress response as well as heat shock proteins and pathogenesis-related proteins are often identified as involved in the stress response while signaling and membrane proteins are frequently underrepresented in the traditional total protein extract 2DE approach. The scope of proteomics studies will therefore have to broaden beyond gel-based approaches to also include high-throughput gel-free approaches as well as specific studies which focus on subproteomes. As more and more crop species are being sequenced, gel-free peptide-based proteomics will probably become the standard for high-throughput crop differential proteomics. Only a few peptide-based abiotic stress studies on crops have been published to date (Ford et al., 2011; Neilson et al., 2011; Mirzaei et al., 2012a; Mirzaei et al., 2012b; Vanderschuren et al., 2013; Buts et al., 2014). Mirzaei and colleagues used a shotgun proteomics approach on rice and provided evidence that protein accumulation patterns are significantly different between moderate and severe drought. They withheld water from rice plants for 14 days, with sampling after 10 days (moderate drought) and 14 days (extreme drought), after which they rewatered the plants. Aquaporins for instance were more abundant in well watered and severely dehydrated plants than in those who were moderately dehydrated or rewatered. A heat shock protein 70 was more abundant under mild drought stress and less abundant in severely stressed plants (Mirzaei et al., 2012a).

2.3.4 Metabolomics

Several metabolic studies have already been conducted in a range of plants. Arbona et al. (2013) and Obata and Fernie (2012) recently reviewed metabolomics research regarding plant abiotic stress. It is not surprising that plants have a diverse metabolic

response towards different abiotic stresses. Obata and Fernie (2012) compared several *Arabidopsis* metabolite studies for different stresses to identify common and stress-specific metabolic responses. Metabolites are accumulated more during abiotic stresses and could potentially be used as building blocks after stress to support a recovery. Levels of sucrose increased during most abiotic stresses. Similarly the levels of osmoprotectants such as raffinose and proline were heavily increased during several stresses. The sugar trehalose on the other hand showed a significantly smaller increase and only in specific stresses and probably has a different function which could be related to the signaling function of its precursor trehalose-6-phosphate. The studied amino acids (valine, isoleucine, leucine, lysine, threonine and methionine) increased more significantly in drought stress than in any of the other stresses and this points to a role as compatible solutes although other roles are also possible. In conclusion a wide array of pathways are regulated under stress and analyses have been performed on carbohydrates, amino acids, polyamines and several secondary metabolites. Many of these metabolites have regulatory, osmoprotective and/or reactive oxygen species scavenging roles. While changes in the primary metabolism show general trends common in several stresses and species, the secondary metabolism is more diversified across species and changes in the secondary metabolism are specific for a particular type of stress (Obata and Fernie, 2012; Arbona et al., 2013).

2.3.5 Phenomics

An overview of phenomics approaches for crops, which can also be integrated in abiotic stress research, was given in section 1.6 of chapter 1. The completely automated phenotyping platforms remain limited to cereals. Further development of high-throughput phenotyping approaches for abiotic stress in crops is certainly necessary as they remain rather sporadic.

2.4 Omics in *Musa*

2.4.1 Genomics and banana

The first progress in the *Musa* genome analysis was made with the construction of BAC libraries by several laboratories in 2003 (Aert et al., 2004; Roux et al., 2008). The first fully sequenced *Musa* genome was published in 2012 (D'Hont et al., 2012). The sequenced genome came from a doubled-haploid Pahang, a *Musa acuminata ssp. malaccensis* genotype. Transposable elements make up about half of the A genome sequence and gene-rich regions are mostly located at the distal parts of the

chromosomes. From the pattern of paralogous gene clusters on the chromosomes, it was inferred that a total of three whole-genome duplications occurred. Most gene-duplicated copies were lost with only 10% retained in four copies and 65.4% are again single copies. When compared to other species, *Musa* specific gene clusters are enriched in transcription factors, defense-related proteins and enzymes of the cell-wall biosynthesis and enzymes of secondary metabolism (D'Hont et al., 2012). In 2013, the publication of a draft *Musa balbisiana* genome followed (Davey et al., 2013). This corresponding B genome was obtained by sequencing a wild diploid *Musa balbisiana* genotype, Pisang Klutuk Wulung, and mapping the reads to the A-genome after which the consensus sequence was extracted. Due to the direct mapping to the A genome instead of independent assembly, the structure and organization of the B genome, especially the positions of transposable elements, cannot be analyzed. Significant sequence divergence from the A genome was observed with one SNP per 39.1 bp. The use of a diploid for the sequencing of the B genome allowed to investigate the heterozygosity within one variety. Heterozygosity accounted for one SNP for every 33.7 bp. Both genomes can be accessed through the Banana Genome Hub which provides several tools to further analyze and use the genomes (Droc et al., 2013).

2.4.2 Transcriptomics and banana

The combination of SuperSAGE and PCR walking by our group resulted in the first analysis of *Musa* gene expression in 2005 (Coemans et al., 2005). A total of 10,000 sequenced tags resulted in 5,292 unique tags, but only half of the 100 most abundant tags could be identified due to limited EST resources. A re-analysis in 2008, when additional EST data were available, resulted in a 76% identification rate for the 50 most abundant tags with an overall matching rate of 36% (Carpentier et al., 2008). A microarray approach was used by Davey et al. (2009) in which *Musa* RNA was hybridized to a Rice GeneChip Genome Array. In total, 2,910 transcripts showed over a two-fold difference in expression levels between control and drought stress treatments. Many differential transcripts, including a number of genes with transcription factor activity, were involved in pathways and processes typically implicated in stress responses. The list of genes also overlapped considerably with earlier findings in the dehydration responses of *Arabidopsis* and rice (Davey et al., 2009). Over the last two years, several groups used RNA-seq technology in banana to assess the transcriptome response to *Fusarium oxysporum* f. sp. *cubense* Tropical Race 4 in both susceptible and resistant cultivars (Li et al., 2012; Bai et al., 2013; Li et al., 2013). These various studies used RNA-seq to assemble a *de novo* transcriptome which was followed by a digital gene expression analysis using short tags that are aligned to the newly assembled transcriptome sequence. Two resistant varieties,

one highly resistant and the other middle resistant, had for instance differentially expressed defense-related genes which could point to different resistance mechanisms. Further research is still needed however to dissect the functions of all genes and the complex interacting pathways (Bai et al., 2013). To our knowledge, no results of abiotic RNA-seq studies have been published in *Musa* so far. Our group is now analyzing RNA-seq data from osmotically stressed leaf and root samples from three different genotypes. Further investigation of several candidates at more time points will be performed using qPCR (Jassmine Zorilla, personal communication).

2.4.3 Proteomics and banana

Our research group has mainly focused on proteomics to research osmotic stress tolerance of *Musa*. Extraction and analysis protocols were optimized for sugar-mediated acclimation research related to cryopreservation (Carpentier et al., 2005; Carpentier et al., 2007; Carpentier et al., 2009; Carpentier et al., 2010). Analysis of the differential proteins revealed that successful sucrose acclimation is probably correlated with an efficient uptake of the sucrose, followed by a reduced breakdown of sucrose to provide an osmoprotective advantage. Moreover, the sugar played an important role as an energy source and in the generation of reducing power as well. Aside from these important changes in proteins related to the sucrose metabolism, several other proteins involved in stress and defense were involved in acclimation as well (Carpentier et al., 2007; Carpentier et al., 2010). The study of membrane proteins is equally important when studying osmotic stress responses. A workflow was set up to identify plasma membrane proteins in *Musa*. The best option was to use a peptide-based approach although identification of peptides and proteins at the time was still severely hampered by the absence of sufficient *Musa* genomic data. Out of 79 identified plasma membrane proteins 19 were predicted to contain at least on transmembrane domain (Vertommen et al., 2011a). Further research now focuses on ABA stress and its influence on the membrane proteome (Suzana Garcia, personal communication).

A *Musa* proteome study from outside our research group investigated the proteome of a *Musa paradisiaca* variety subjected to cold stress using 2D-LC MS/MS and iTRAQ (Yang et al., 2012). The authors first constructed a RNA-seq database and this resulted in a total of 43,313 annotated contigs which enabled protein identification. The proteomics approach quantified 2658 proteins which resulted in 809 unique proteins with differential abundances. These proteins predominantly belonged to stress response, primary metabolic and oxido-reduction pathways. By comparing the cold stress response in a sensitive banana variety with response in the more tolerant *Musa paradisiaca* variety through Western blot and enzymes activity assays, they concluded that more effective reactive oxygen species scavenging through the

catalase pathway might partly explain the greater tolerance of the *Musa paradisiaca* variety to cold (Yang et al., 2012). Proteome analyses of the root response to *Fusarium oxysporum* f. sp. *cubense* Tropical race 4 revealed the importance of many proteins related to the defense pathways, including pathogenesis-related proteins, signal conduction proteins, molecular chaperones and oxidative-redox homeostasis proteins (Li et al., 2012). Susceptible and resistant varieties showed different responses, which is likely linked with their different resistance levels, but further research is needed. To search for low abundant proteins in banana fruit, Esteve and colleagues used a beads-based combinatorial peptide ligand library approach (Esteve et al., 2013). A total of 1131 proteins were identified. The captures with the beads led to the identification of 849 proteins while untreated samples led to the identification of 452 proteins with 170 proteins in common between both methods. They identified several of the known *Musa* allergens as well as proteins related to starch degradation and involved in fruit ripening.

2.4.4 Metabolomics and banana

Metabolomics studies in banana are mostly limited to biotic stress responses. Recent research showed that a phenalenone-type phytoalexin, a secondary metabolite, mediates banana resistance towards a burrowing nematode (Hölscher et al., 2013). The compound was present in higher concentrations in lesions of the resistant cultivar and an *in vitro* bioassay confirmed its nematostatic and nematocidal effect. The formation of similar compounds was also identified in an earlier study as induced in roots by *Sporobolomyces salmonicolor* and other phenalenone-type compounds showed activity against *Mycosphaerella fijiensis* in a bio-assay (Otálvaro et al., 2007; Jitsaeng and Schneider, 2010).

An abiotic stress metabolite study on banana meristems by our group focuses on sugar, sterol and fatty acid composition caused by sucrose-induced acclimation (Zhu et al., 2006). We observed that in untreated meristems sucrose and total sugar content were linked to post-thaw recovery suggesting that these two factors are crucial for survival after cryopreservation. On the other hand, the accumulation of sugars in sucrose pretreated meristems could not explain the variability in survival rates between varieties. This suggests that a minimal amount of sugar is needed to survive cryopreservation. It was also observed that varieties in which membrane changes were minimal were the best survivors after cryopreservation (Zhu et al., 2006).

2.4.5 Phenomics and banana

As shown in Ravi et al. (2013), abiotic stress phenotyping in banana is still mainly based on manual measurements. At the laboratory of Tropical Crop Production, however, several systems have been developed and are under development to monitor growth and transpiration responses to stress. Leaf area calculation software was developed to automatically analyze the leaf area of greenhouse plants based on RGB photography (Ewaut Kissel, personal communication). This set-up is not yet automated due to the size and weight of banana plants and the limited amount of extra space for a robot in the greenhouse, but plans are being developed to automate weighing and irrigation. These data are used for growth monitoring, transpiration calculations and to determine the amount of irrigation needed per plant. The use of a robot could increase the number of measuring and watering time points to several times a day instead of two or three times per week. An infrared camera in combination with the RGB leaf area calculator is used on autotrophic *in vitro* plants, grown in controlled climate chambers, to assess leaf temperature. A set-up is now being developed to continuously monitor transpiration of the plants in these growth chambers.

II. Experimental results

Chapter 3

Evaluating potential osmotic stress markers using qPCR

Proteomics data were generated by Sebastien Carpentier (SC) and EST data were generated in collaboration with EMBRAPA. SC and A-CV selected the four evaluated stress markers. A-CV designed the primers, performed the qPCR experiment and analyzed the data.

3.1 Introduction

The global *Musa* Germplasm Collection is stored at Bioversity's International Transit Centre at KU Leuven. Over 1400 accessions are maintained *in vitro* and more than 850 accessions have been cryopreserved as well. The cryopreservation process consists of several steps: meristem cultures are first subjected to an osmotic acclimation treatment for two weeks, afterwards they are severely dehydrated and quickly frozen in liquid nitrogen. The sucrose-mediated osmotic acclimation step has been shown to be crucial to the post-thaw survival and regeneration of *Musa* (Panis et al., 2002). Further research investigated the mechanisms behind this sucrose-mediated acclimation using proteomics and transcriptomics (Carpentier et al., 2007; Carpentier et al., 2010). Sucrose however is not purely an osmotic stressor as it is the main end-product of photosynthesis and the most translocated sugar. Furthermore, sucrose is an important signaling molecule during normal development and during stress (Rolland et al., 2002). Therefore, our acclimation research also included sorbitol as a non-metabolized sugar to study only the effects of the osmotic stressor without the effect of the carbon source and/or signaling component. 2D-DIGE and MALDI-TOF/TOF MS were used and a genotype-specific EST library was generated (Carpentier et al., 2010).

The EST database was used as an additional source for protein identifications aside from cross-species identification. A total of 11070 reads were assembled into 1433 contigs. This resulted in 75 additional identifications or 12% of the total identified spots (Carpentier et al., 2010). More than fifty proteins/genes were identified as potential stress markers in the proteomics experiments.

The EST libraries were also used to estimate transcript levels of control versus stress treatments (0.09 M or 0.4 M sucrose for 2 weeks). The number of ESTs belonging to a certain contig were counted and compared with each other. The EST database covered genes that code for proteins that are too big or too small, too low abundant, outside of the pI range or were too hydrophobic, such as membrane proteins, and were consequently not identified on 2DE gels.

This combined 2DE and EST approach resulted in the identification of more than fifty potential stress markers. To evaluate whether they are true stress markers and to specify whether they are osmotic stress markers, we used qPCR to follow the transcription levels over time. We selected four promising potential osmotic stress markers: pathogenesis-related protein 10 (PR10), SUMO-conjugating enzyme, an ABA-responsive protein and phosphoglycerate kinase. The qPCR experiment was performed on the Cachaco variety which is known to have a high survival rate after cryopreservation (Panis et al., 2002). It was shown that the optimal shoot

regeneration for the Cachaco variety was obtained with a four-day acclimatization period prior to cryopreservation (Carpentier et al., 2010). Therefore we selected time points within the optimal four-day acclimation period (0, 2, 12, 24 and 96 hours).

3.2 Experimental procedures

3.2.1 Selection of the four potential stress markers

Using the experimental data from earlier 2DE, EST and qPCR approaches, four interesting potential stress markers were selected based on their significant differential expression in the protein and qPCR experiments or their overrepresentation in the stress EST library compared to the control EST library (Carpentier et al., 2007; Carpentier et al., 2010; Henry et al., 2011).

3.2.2 Plant material

In vitro plants of the selected variety Cachaco (ABB, ITC 0643) were supplied by the International Transit Centre of Bioversity International. Multiple shoot meristem cultures were initiated as described by Strosse *et al.* (Strosse et al., 2006) and maintained on a standard control medium (MS medium supplemented with benzylaminopurine (Duchefa Biochemie, Netherlands)).

Meristem cultures were cut to similar size (approximately 8 mm x 8 mm x 3 mm) and transferred to either fresh standard medium containing 0.09 M sucrose (control treatment) or fresh medium containing an additional 0.21 M sorbitol (stress treatment). A control set was frozen in liquid nitrogen immediately after cutting (0 hours). Meristems were harvested after 2, 12, 24 and 96 hours and flash frozen in liquid nitrogen. Meristems were collected and divided in six biological replicates for each treatment and stored at -80°C.

3.2.3 Total RNA extraction

Material was ground in liquid nitrogen for RNA extraction. Total RNA was extracted using the RNeasy Plus Mini kit (Qiagen, Germany), according to the manufacturer's recommendations except for the addition of PVP40000 to the extraction buffer to a final concentration of 5 mg/ml. The extracted RNA was treated with DNaseI (AB Applied Biosystems, Belgium) for 45 min at 37 °C and finally purified using a phenol-chloroform-isoamyl alcohol extraction/ethanol precipitation purification step. The quality and quantity of the RNA ($A_{260/230}$ and $A_{280/260}$) was determined using the

Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies). Only samples with ratios above 1.8 were used for further analysis. To confirm the absence of genomic DNA after the DNase treatment, a real-time PCR was performed on the treated RNA using the EF1α primers using the conditions below. Only samples for which no amplification could be detected after 40 cycles, were used for analysis. Samples which were still contaminated were treated again with DNaseI and checked once more. A maximum of three DNase treatments per sample was performed.

3.2.4 Primer design

Transcript levels of a set of five reference genes (actin, tubulin, elongation factor-1, 25S r-DNA and ribosomal protein L2) were analyzed as well as those of the candidates PR10, SUMO-conjugating enzyme, phosphoglycerate kinase and ABA-responsive protein. The primers for the reference genes, ABA-responsive protein and phosphoglycerate kinase had already been designed and optimized for use (Henry et al., 2011; Podevin et al., 2012). For PR10 and SUMO-conjugating enzyme, primers were designed based on the EST sequences available at that time (2010). Primer pairs were designed using the Primer3 software[11] with the following parameters: a length of 19-25 bp, a melting temperature of 58-60 °C, GC content of 45-60 %, a maximum (self) complementarity of 4, a maximum 3' (self) complementarity of 1 and amplicon size of 75-200 bp. Primers were further analyzed for hairpin, self-dimer and heterodimer formation using OligoAnalyzer 3.1[12]. The primer pairs were tested by gradient PCR on cDNA and gDNA to check for specificity and optimal annealing temperature. The primer pairs for qPCR analysis of the genes of interest with their optimal annealing temperature are presented in Table 3.1.

Table 3.1: List of primers for the stress markers

Gene	Primer name	Sequence	Annealing temp (°C)	Amplicon length (bp)
PR10	PR10-F	CATGTTGCTGCCATCTCTCT	62	76
	PR10-R	TCCTTAGACGACCACACAAAAC		
SUMO-conjugating enzyme	SCE-F	TCCCTCTTACTGTCCATTTCAG	62	141
	SCE-R	GTCTCCATCCACTGTCTTCATT		
Phospoglycerate kinase	PGK-F	ATCATCGGAGGTGGTGACTC	60	147
	PGK-R	TTAGGCATCTTCAAGAGCAAG		
ABA-responsive protein	ARP-F	GCTTGCTACCTCTCGACCAC	60	129
	ARP-R	GTAGCTCCAGGCTTGCTGAC		

[11] http://frodo.wi.mit.edu/primer3/

[12] http://eu.idtdna.com/analyzer/applications/oligoanalyzer/

3.2.5 Two-step real-time RT-PCR

One microgram of RNA was reverse transcribed into cDNA using the RevertAid H Minus First Strand cDNA synthesis kit (Fermentas, Germany) and oligo(dT)$_{18}$ primers according to the manufacturer's instructions.

The qPCR was performed in a Corbett Rotor-Gene 3000 (Qiagen, Germany). The reactions included 1x ABsolute qPCR SYBR Green I mix (Thermo Scientific, United Kingdom), 150 nm of reverse and forward primers, 2 µL of template (50x diluted cDNA, gDNA, λ-DNA or water) and water to a total volume of 25µL. Each run contained all 0 hour samples (6 replicates), two replicates of all other treatments, a standard curve of six serial four-fold dilutions of pooled cDNA and non-template control samples. Two technical replicates were run per sample and averaged for analysis. This resulted in a total of 3 runs to run all samples per gene. The following amplification program was used: 15 min at 95 °C followed by 45-50 cycles of 15 s at 95 °C, 20 s at 52-62 °C (depending on the annealing temperature for the gene being run, see Table 3.1 and Podevin et al. (2012)), 30 s at 72 °C and a final fluorescence measurement step at 79-81 °C for 15 s. At the end of each qPCR run a melting curve was produced from 55 °C to 93 °C to verify the specificity of the amplicon for each primer pair.

3.2.6 qPCR data analysis

Cq (quantification cycle) values were converted to relative quantities using the gene-specific PCR efficiency calculated from a standard dilution series (Hellemans et al., 2007). All samples were included to calculate the reference gene stability measure M using geNorm v3.5 software (Vandesompele et al., 2002). The three reference genes used for normalization were actin, tubulin and elongation factor 1 (see section 3.3.1). Afterwards an inter-run calibration was performed per gene using the geometric mean of the 0 hours control samples which were included in each run (Hellemans et al., 2007). These calibrated normalized relative quantities were converted to relative mRNA abundances as fold changes compared to the lowest measured expression for that gene. Statistical analysis of the relative mRNA abundances was performed using the non-parametric Kruskal-Wallis test using R v3 and the Agricolae package. Box plots were generated using STATISTICA v10 software.

3.3 Results and Discussion

3.3.1 Reference gene selection

The geNorm software was used to select the most stably expressed reference genes (Vandesompele et al., 2002). The average expression stability values (M) of remaining control genes during stepwise exclusion of the least stable control gene are represented in Figure 3.1a. It shows a ranking of the candidate reference genes according to their expression stability with the most unstably expressed genes which have a high M value at the left and the best reference genes with a low M value at the right. A minimum of three reference genes is recommended for normalization and more genes can be included if this is necessary for stability. To determine the optimal number of reference genes the pairwise variation ($V_{n/n+1}$) between two sequential normalization factors containing an increasing number of less stable reference genes, NF_n and NF_{n+1}, is calculated (Figure 3.1b). V3/4 for instance compares the use of the three most stable reference genes (actin, tubulin and elongation factor 1 which have the lowest M values) versus the four most stable reference genes (actin, tubulin, elongation factor 1 and ribosomal protein L2) when calculating the normalization factor. A small variation (cut-off value < 0.15) means that the added reference gene does not have a big effect on the normalization factor and therefore the extra reference gene does not have to be included. Based on the V3/4 value, there is no need to include a fourth or fifth reference gene and the three

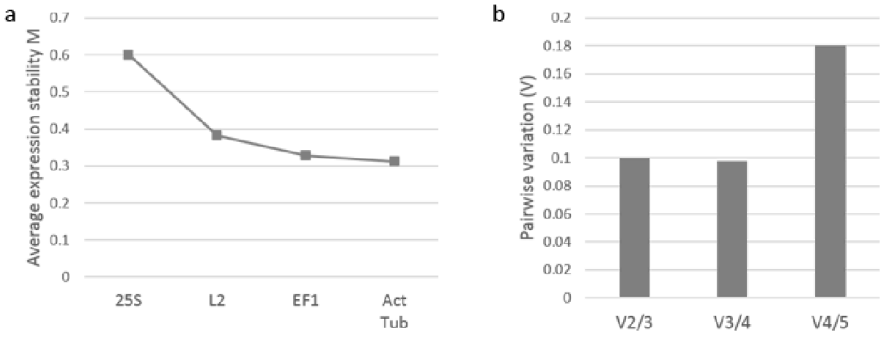

Figure 3.1: Expression stability and variation analysis of the candidate reference genes. ***a*** *Average expression stability (M) of remaining control genes and ranking of the candidate reference genes during stepwise exclusion of the least stable control gene. 25S: 25S r-DNA, L2: ribosomal protein L2, EF1: elongation factor-1, Act: actin and Tub: tubulin.* ***b*** *Pairwise variation ($V_{n/n+1}$) analysis between the normalization factors NF_n and NF_{n+1} to determine optimal number of reference genes (cut-off value 0.15).*

most stable reference genes are sufficient. Even the V2/3 is already below the cut-off value which means that the normalization factors calculated with the three reference genes are not very different from those calculated on two reference genes and that two reference genes would already provide a relatively good normalization. The three most stable reference genes, actin, tubulin and elongation factor-1, were therefore selected to carry out the normalization. These reference genes had also been shown in other experiments using meristems to be the most stable (Podevin et al., 2012).

3.3.2 Evaluation of the potential stress marker genes

PR10 is the first potential stress marker as a PR10 protein spot was more abundant under stress than during control conditions (Carpentier et al., unpublished results). We also identified a contig showing homology to PR10 in the Cachaco *Musa* EST database. Further analysis of the contig showed that it contains the Bet v I domain (CDD analysis score: 2.98e-43)[13], a known subclass of the PR10 family. The pathogenesis-related proteins are actually a collection of unrelated proteins that are all involved in the defense system. They are divided into 17 classes (PR1-PR17) based on sequence homology and similar biological activity or physicochemical properties (van Loon and Van Strien, 1999; van Loon et al., 2006). While PR10s are known for their expression during both biotic and abiotic stresses, some members are expressed constitutively. The latter indicates that aside from their protective role they might also have a more general role in plant development. Yet their actual functions, even during stress, still remain unknown (Fernandes et al., 2013). The most well-known subclass of constitutively expressed PR10 is undoubtedly those constituting a large group of food and pollen allergens in birch pollen, apple, celery and other fruits and vegetables (Liu and Ekramoddoullah, 2006) to which the *Musa* PR10 also shows homology.

Our qPCR analysis shows that the expression of the PR10 gene after 12 hours and 24 hours was significantly higher than after 0 days and 4 days in both control and stress conditions (Figure 3.2). Only at 12 hours, the PR10 gene expression was significantly higher under stressed than under control conditions. Based on these observations, the applied osmotic stress does not seem to be the only inducer of the gene expression for PR10 in short term acclimation stress. The upregulation of expression in both control and stress conditions is probably due to the cutting and transfer of the meristem cultures to a new medium at the start of the experiment rather than

[13] http://www.ncbi.nlm.nih.gov/cdd/

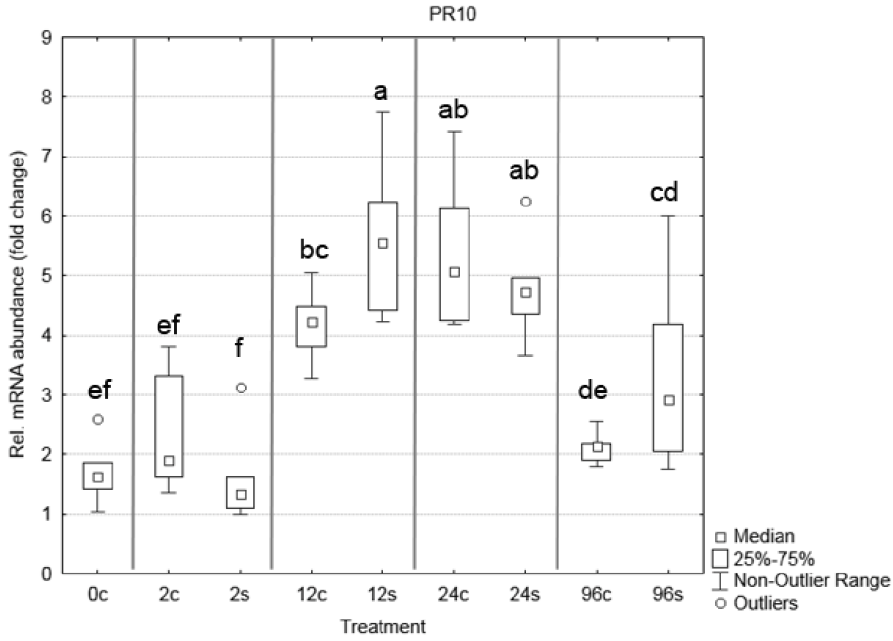

Figure 3.2: Relative mRNA abundance of PR10 relative to the lowest expression of the gene following exposure to sorbitol stress (0.21 M). Banana meristems were subjected to a control (c) or stress (s) treatment for 0, 2, 12, 24 or 96 hours. Bars marked with the same letter do not significantly differ from each other, a>b>c>d>e>f (α 0.05, n = 6, outliers= 1.5x interquartile range).

the application of the osmotic stress. This effect seems to have passed after 96 hours as control levels at 96 hours resemble the 0 hours control levels.

The second potential stress marker, SUMO-conjugating enzyme, was upregulated in the *Musa* EST library under stress conditions. A total of 28 ESTs belonged to the contig in the stress library versus 1 EST in the control library. Although ubiquitination remains the most well-known post-translational modification of proteins by a small polypeptide, the small ubiquitin-like modifier (SUMO) has also been implicated in many abiotic stress responses over the last decade as reviewed by Castro et al. (2012). Sumoylation and ubiquitination are very similar multi-step processes mediated by E1 activating enzymes, E2 conjugating enzymes and E3 ligase. The enzymes used in the two pathways are similar but have specific structural features (Downes and Vierstra, 2005). SUMO-conjugating enzyme is encoded by a single gene in *Arabidopsis thaliana* but a phylogenetic analysis has shown that tomato, grapevine, poplar, rice, *Brachypodium*, sorghum and maize all encode two or more SUMO-conjugating enzyme genes (Novatchkova et al., 2012). It was observed that the monocots in this study had additional SUMO-conjugating enzyme genes with a slightly different sequence which might be considered a monocot-specific subgroup.

SUMO-conjugating enzyme expression was induced after 2 hours and reached a maximum at 12 hours after the start of the experiment and baseline levels were approached again after 4 days (Figure 3.3). Both control and stress samples showed the same expression profile except after 24 hours when a lower quantity was measured in the stress samples than in the controls. These results differ from earlier observations in which SUMO-conjugating enzyme ESTs from stress samples after 14 days on 0.4 M sucrose were more abundant than in control conditions at that same time point (28 ESTs versus 1). One has to take into account however that the sucrose stress was more severe than the sorbitol stress applied in this experiment (water potential of -1.074 MPa with 0.4 M sucrose versus -0.756 MPa with 0.09M sucrose and 0.21 M sorbitol) and that the stress was applied for 14 days instead of a maximum of 4 days in our qPCR experiment and that sucrose has a possible signaling effect. The similar induction of the transcript level, which starts after two hours, in both control and stressed meristem cultures in our qPCR analysis suggests that the wounding and transfer to new medium affect the SUMO-conjugating enzyme

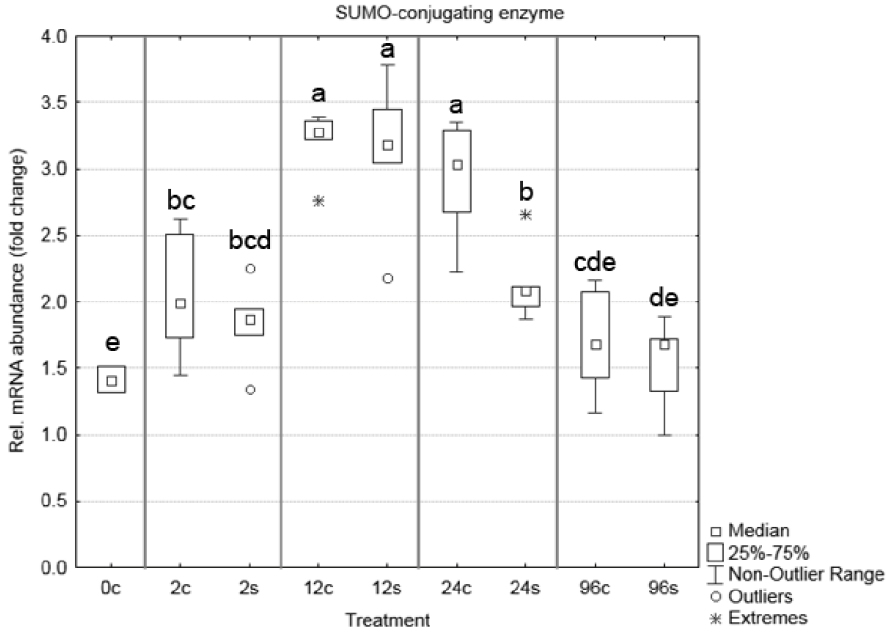

Figure 3.3: Relative mRNA abundance of SUMO-conjugating enzyme relative to lowest expression of the gene following exposure to sorbitol stress (0.21 M). Banana meristems were subjected to a control (c) or stress (s) treatment for 0, 2, 12, 24 or 96 hours. Bars marked with the same letter do not significantly differ from each other, a>b>c>d>e (α 0.05, n = 6, outliers = 1.5x interquartile range, extremes = 3x interquartile range).

transcript level more than the osmotic stress. The effect of this wounding and transfer stress seems to disappear however after 96 hours as control treatments at 0 hours and 96 hours show a similar level of mRNA abundance.

In earlier qPCR experiments, ABA-responsive protein transcript levels were shown to be highly upregulated after ABA treatment in meristem cultures (Henry et al., 2011). Since the plant hormone ABA probably also plays a signaling role in osmotic stress, we selected the ABA-responsive protein as a potential stress marker. Primers weere designed based on ESTs with homology to the *A. thaliana* 'ABA-responsive protein-like' gene (At5g13200) (Henry et al., 2011). This gene was shown to be an ABA-induced gene in both *A. thaliana* and rice (Hoth et al., 2002; Yazaki et al., 2004). TAIR, a database of genetic and molecular biology data for *Arabidopsis thaliana*, now annotates it as a GRAM domain family protein (from *G*lucosyltransferases, *Ra*b-like GTPase activators and b-like GTPase activators and *M*yotubularins). The GRAM domain is probably involved in membrane-associated processes such as several signaling pathways. Little is known about the actual function of the different members of this family, but a high expression divergence in different tissues and to different stresses was observed in both *Arabidopsis* and rice (Jiang et al., 2008). That study also showed no significant response of the At5g13200 to ABA, however, a significant upregulation to osmotic (PEG) and salt stress was observed.

The expression of the ABA-responsive protein gene was significantly upregulated after two hours but returned to normal levels after 12 hours (Figure 3.4). Since both control and sorbitol stress levels were similarly upregulated after two hours, we hypothesize that the wounding and transfer to the new medium rather than the sorbitol stress is responsible for its induction. At all other time points a small but statistically significant upregulation in stress treatments was observed. The applied osmotic stress therefore does have an effect on the expression of ABA-responsive protein but this effect is much smaller than the initial wounding and transfer effect.

The final stress marker candidate is phosphoglycerate kinase. Six phosphoglycerate kinase spots were already identified in the 2007 proteomics study and four of these spots were more abundant under sucrose stress (Carpentier et al., 2007). In a 2010 study it was shown that phosphoglycerate kinase reacted the most to sucrose stress but that phosphoglycerate kinase abundance was also significantly higher during sorbitol stress than in control conditions (Carpentier et al., 2010). Earlier qPCR experiments confirmed an increased mRNA expression in sucrose-treated meristems (Henry et al., 2011). The upregulation of phosphoglycerate kinase fits with the hypothesis of tolerance as it is a glycolysis enzyme that catalyzes the formation of 3-phosphoglycerate from 1,3-bisphosphoglycerate while also one ATP is produced

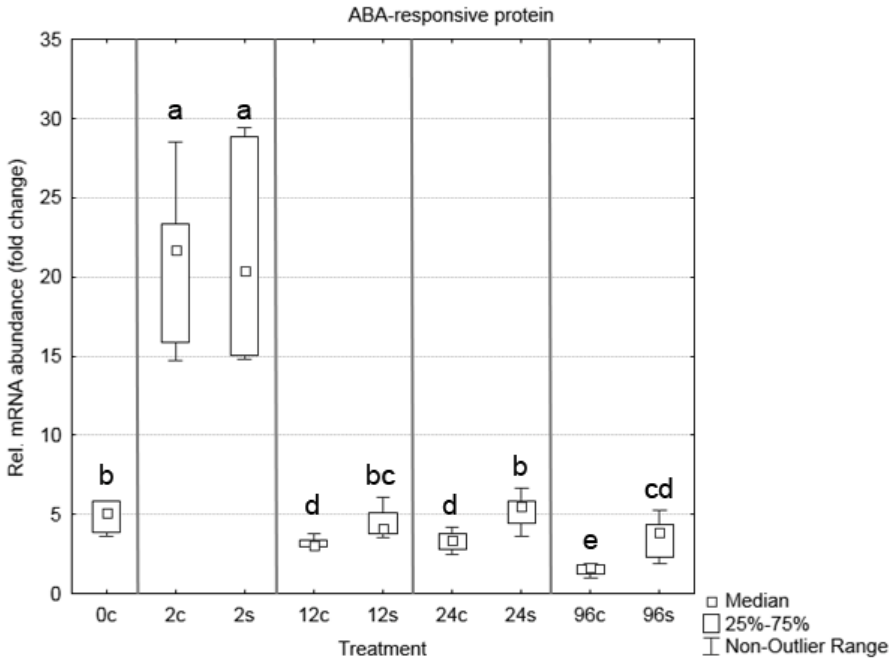

Figure 3.4: Relative mRNA abundance of ABA-responsive protein relative to the lowest expression of the gene following exposure to sorbitol stress (0.21 M). Banana meristems were subjected to a control (c) or stress (s) treatment for 0, 2, 12, 24 or 96 hours. Bars marked with the same letter do not significantly differ from each other, a>b>c>d>e (α 0.05, n = 6).

which generates the necessary energy for other stress response mechanisms (Carpentier et al., 2010).

The control condition only showed a slight increase in phosphoglycerate kinase transcript level over time. The sorbitol-stressed samples on the other hand showed a significant increase in expression after 12 hours which is already significantly lower again after 4 days (Figure 3.5). Earlier qPCR experiments had shown that phosphoglycerate kinase expression was also upregulated by high sucrose treatment (Henry et al., 2011). So osmotic stress treatments, both the metabolized sucrose and the non-metabolized sorbitol, upregulate phosphoglycerate kinase expression. Transcript level increases of phosphoglycerate kinase seem to be in line with the higher phosphoglycerate kinase abundance observed in proteomics experiments (Carpentier et al., 2010). The earlier qPCR experiments also showed that phosphoglycerate kinase expression was not affected by cutting (Henry et al., 2011). This is corroborated by our results as we only observe a slight increase in transcript

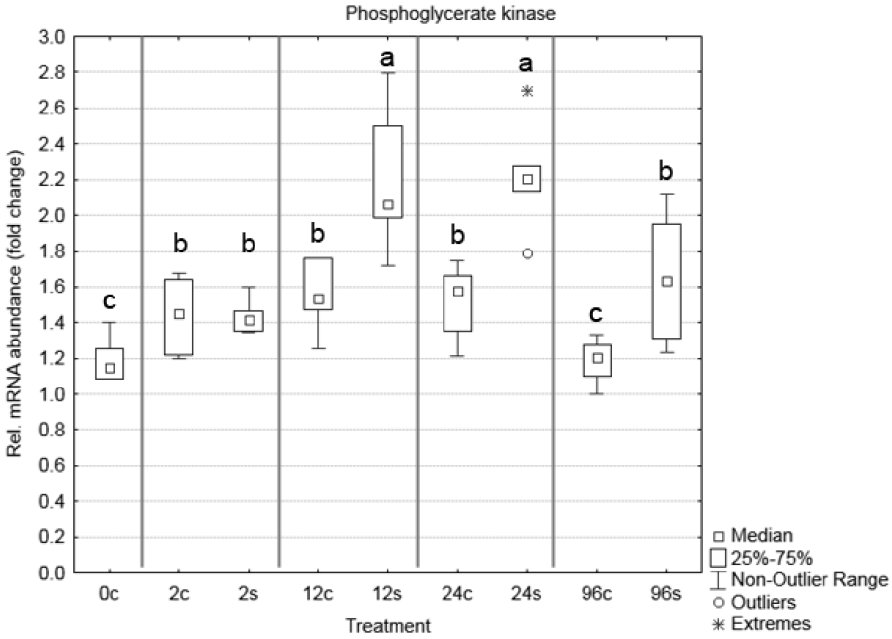

Figure 3.5: Relative mRNA abundance of phosphoglycerate kinase relative to the lowest expression of the gene following exposure to sorbitol stress (0.21 M). Banana meristems were subjected to a control (c) or stress (s) treatment for 0, 2, 12, 24 or 96 hours. Bars marked with the same letter do not significantly differ from each other, a>b>c (α 0.05, n = 6, outliers = 1.5x interquartile range, extremes = 3x interquartile range).

levels in the control which could be related to the supply of sucrose in the new medium. We can therefore conclude that phosphoglycerate kinase is a suitable osmotic stress marker which does not react to wounding.

The primary goal of our experiments was to find suitable osmotic stress marker genes. All of the genes did respond to the stress mediated by wounding and/or transfer to new medium and phosphoglycerate kinase seems the most suitable marker for osmotic stress as it is barely affected by wounding and only slightly by transfer to a new medium. While PR10, SUMO-conjugating enzyme and ABA-responsive protein react to wounding and transfer more than to sorbitol stress in this short-term experiment, they are good stress markers towards the treatment the meristems are subjected to before cryopreservation. SUMO-conjugating enzyme did not seem to respond to the sorbitol treatment at all, but ABA-responsive protein and PR10 transcript levels did show some response to the osmotic stress treatment.

The wounding and transfer effect on mRNA abundance seems to last less than 96 hours as control treatment levels at 96 hours have returned to the levels measured

at the start of the experiment. Based on the four stress markers we conclude that the meristem cultures are already stressed after two hours and acclimated to wounding after 96 hours which corresponds with the observed survival after cryopreservation earlier (Carpentier et al., 2010).

3.4 Conclusions

All stress markers genes can be used to detect short-term stress in meristems. Although we can conclude that wounding and transfer to a new medium seem to have the greatest short-term effect on the expression of most of the tested candidate genes, the phosphoglycerate kinase gene is a good candidate to be used as an osmotic stress marker in meristems.

Chapter 4

Screening the banana biodiversity for drought tolerance: can an *in vitro* growth model and proteomics be used as a tool to discover tolerant varieties and understand homeostasis

This chapter is based on the manuscript:
Screening the banana biodiversity for drought tolerance: can an *in vitro* growth model and proteomics be used as a tool to discover tolerant varieties and understand homeostasis
Anne-Catherine Vanhove, Wesley Vermaelen, Bart Panis, Rony Swennen and Sebastien Christian Carpentier
Frontiers in Plant Science, doi:10.3389/fpls.2012.00176

Yves Lambeens (YL) performed the *in vitro* growth test and the proteomics experiment. A-CV analyzed the growth and proteomics data. Protein identification was performed by SYBIOMA. A-CV and SC wrote the manuscript.

4.1 Introduction

There is a great need for research aimed at understanding drought tolerance, screening for drought tolerant varieties and breeding crops with an improved water use efficiency. Drought is one of the major abiotic stress factors in most crops lowering yields considerably. Agriculture currently uses 70% of water withdrawn worldwide but demands in water are still rising. Climate change and an increasing world population will result in even more water needed for food production but demands will also rise in the municipal and industrial sector (WWAP, 2012). To meet the demands of the future world, crops will need to be produced more efficiently, meaning agriculture needs to produce 'more crop per drop'.

Bananas and plantains are a major staple food and export product in many countries with a worldwide production of over 135 million tonnes per year (FAO, 2012). Even though bananas are only grown in the humid tropics and subtropics, in many locations rainfall is not sufficient or not evenly distributed throughout the year. Commercial plantations supplement this rainfall with irrigation, but for small farm holders this is not feasible. Water is one of the most limiting abiotic stress factors in banana production. Bananas need at least 25 mm of water per week and an annual rainfall of 2000-2500 mm evenly distributed along the year is considered optimal for banana production. When there is no access to irrigation, mild drought conditions are responsible for considerable yield losses. Van Asten et al. (2011) calculated a yield loss of up to 65% when the annual rain fall was below 1100mm - still an enormous amount of precipitation. Moreover in the humid tropics bananas are threatened by the disease Black Sigatoka, caused by *Mycosphaerella fijiensis*. Export bananas, all from the Cavendish subgroup, are extremely susceptible and economic damages rise due to yield loss and the cost of the chemical inputs that are required to control the disease. Cultivating bananas in drier areas where the infection rate is much lower, would be an alternative (Marin et al., 2003; Robinson and Sauco, 2010).

Cultivated banana varieties are hybrids of two wild diploid species *Musa acuminata* (genome constitution AA) and *Musa balbisiana* (genome constitution BB). Most cultivated varieties are triploids with either an AAA, AAB or ABB genome constitution. Varieties with an AAB or ABB genome constitution are said to be more drought tolerant and hardy due to the presence of the B genome (Simmonds, 1966; Thomas et al., 1998; Robinson and Sauco, 2010). The commercially exploited varieties are triploids with an AAA genome constitution which are sweet and extremely suitable to harvest immature, transport and ripen upon arrival. However, this AAA Cavendish group is drought sensitive. We at KU Leuven host Bioversity's International Transit Centre that contains the *Musa* International Germplasm

collection with over 1400 accessions and we want to explore this biodiversity for tolerant varieties.

While survival mechanisms, such as closing stomata, reducing leaf area and growth arrest under drought conditions is a good survival mechanism for plants in the wild, from an agricultural point of view growth reduction only lowers yield. A growth stop or a serious growth reduction when the drought stress is non-lethal is unwanted. Experiments under severe stress conditions tend to select slow growing varieties that are able to survive a long period of severe drought. But those conditions are seldom applicable to agricultural conditions and certainly not to banana. It has also been indicated that severe stress conditions activate different mechanisms that are not necessarily relevant to agricultural conditions (Skirycz et al., 2011). We are looking for vigorous plants that will only show a minor reduction in growth, photosynthesis and metabolism under mild drought or osmotic stress. Acclimation to mild stress will require a new homeostasis so that the plant can continue growing during stress.

Many plant collections are kept as seeds or in the case of banana as *in vitro* plantlets. The most straightforward way to characterize and screen an *in vitro* collection is to immediately evaluate the *in vitro* plantlets. So the first logical step to screen the *Musa* biodiversity for possible drought tolerant varieties was the development of a suitable *in vitro* test (Rukundo et al., 2012). Shekhawat and colleagues report a similar *in vitro* test to evaluate the osmotic tolerance of a transgenic banana (Shekhawat et al., 2011b). However how relevant is an *in vitro* growth model towards field conditions? We designed a long term experimental setup to check this as discussed in the Rationale and outline of this PhD research. The advantages of this first *in vitro* model to screen the *Musa* biodiversity are the throughput and the possibility to control the experiment; the disadvantages are the artificial conditions.

Abiotic stress research in *Musa* is still in its infancy. Some valuable research has been done in the past by several groups (Carpentier et al., 2007; Fan et al., 2007; Carpentier et al., 2010; Liu et al., 2010; Henry et al., 2011; Shekhawat et al., 2011a; Shekhawat et al., 2011b). In this study we present the results of a selection for tolerant varieties using the optimized *in vitro* model and the proteome analysis of the most tolerant variety.

4.2 Experimental procedures

4.2.1 Heterotrophic *in vitro* test

In vitro plants were supplied by the Bioversity International *Musa* Germplasm collection. The selected varieties were the highland (h) variety Mbwazirume (AAAh, ITC 0084), the Cavendish variety Williams (AAA, ITC 0365), Popoulou (AAB, ITC 0335), the plantain (p) variety Obino L'Ewai (AABp, ITC 0109) and Cachaco (ABB, ITC 0643). Plants were multiplied on semisolid p5 medium consisting of Murashige and Skoog basal salts and vitamins supplemented with 10 µM benzylaminopurine, 1 µM indole acetic acid, 10 mg/l ascorbic acid, 0.09 M sucrose and 3 g/l Gelrite® (Strosse et al., 2006). Experiments were carried out on a liquid p6 medium, the same as p5 but with 1 µM benzylaminopurine and without Gelrite®: (i) standard control medium (containing 0.09 M sucrose) and (ii) stress medium containing 0.09 M sucrose and 0.21 M sorbitol. Well-developed plantlets were excised from multiple shoot clusters from the p5 medium and put on liquid p6 medium. After 4 weeks all leafs were removed and explants of about 3 cm of length with three roots of about 1 cm were excised. The explants were then put on the control or stress medium for 48 days. Medium was refreshed every two weeks. The plants were weighed at the beginning and end of the experiment and the total growth was calculated. Statistical analysis was performed using STATISTICA software 10. At day 48, leaf samples were frozen in liquid nitrogen and stored at -80 °C for protein extraction.

4.2.2 Proteomics

Leaf proteins from 6 control and 6 stressed plants were extracted using the phenol extraction/ammonium acetate precipitation protocol described by Carpentier et al. (2005). DIGE labeling of the protein samples with CyDyes was performed (GE Healthcare). The internal standard, obtained by pooling equal amounts of all protein samples was labeled with Cy2. Control and stresses samples were evenly distributed over Cy 3 and Cy5 and a control and stress sample were included in each gel. The labeled samples were pooled, separated on gel and scanned according to Carpentier et al. (2009). Data were analyzed using the DeCyder software 7.0 (GE Healthcare). The estimated number of spots was set at 10,000 and an exclusion filter based on volume was used to eliminate the detected non-proteinaceous spots as recommended. For spot picking, the proteins were visualized using a colloidal G250 Coomassie Brilliant Blue staining (Neuhoff et al., 1988) after the scanning of the fluorescent dyes. Gel pieces were treated as described by Shevchenko et al. (2006). The samples were resuspended in Milli-Q (MQ) water containing 5% acetonitrile (ACN) and 0.1% formic acid (FA) and separated on an HPLC system, equipped with a

C18 precolumn (PepMap 100, 5 µm – 100 Å, 0.3 x 5 mm, Dionex) to concentrate and desalt the sample. After loading the sample, the following gradient was applied for the mobile phase: solvent A (99.9% MQ / 0.1% FA), solvent B (99.9% ACN / 0.1% FA), from 5% B to 20% B in 2 minutes, to 35% B in 8 minutes, to 45% B in 4 minutes to finally in 95% B in 1 minute, at a flow rate of 250 nL/min over the analytic column (Pepmap 100, 3 µm – 100 Å, 75 µm x 5 cm, Dionex). After LC separation, peptides were positively ionized at 1.7 kV, at 200 °C and injected into the mass spectrometer. Mass spectrometry data were acquired in a ProteomeX-LTQ Workstation (Thermo Scientific, USA) in data-dependent acquisition (DDA) mode controlled by Xcalibur 1.4 software (Thermo Scientific). The typical DDA cycle consisted of a full scan within m/z 400 to 1,600 range followed by five separate data-dependent scans, each taking the 1st to 5th highest peak respectably, under normalized collision energy of 35%. Fragmented precursor ions were dynamically excluded according to the following: repeat counts: 2, repeat duration: 15 s, exclusion duration: 180 s. Peak detection and convertion to 'mgf'-files was performed using MS Convert from ProteoWizard 3.0.3631 software, with the following filter: ChargeStatePredictor 4 1 0.9. Two database searches were performed using an in house Mascot server version 2.2.04 against the NCBI Viridiplantae database (852,488 sequences) and against an in house database that is constructed from all the *Musa* proteins known in NCBI complemented with EST data from different experiments (Carpentier et al., 2008; Carpentier et al., 2010) and the sequences of trypsin and keratin resulting in a concatenated search database containing 169,829 unique entries. Estimation for false positives was made by searching in Mascot against the equivalent decoy database. Search parameters were set as follows: oxidation of methionine was allowed as a variable modification and carbamidomethylation of cysteine as a static modification; enzyme: trypsin; number of allowed missed cleavages: 1; peptide tolerance: 1,000 ppm; fragment ions tolerance: 1.2 Da, instrument type: ESI-TRAP. Results of both searches were exported as csv files and combined in one pivot table (Microsoft Excel). The significant protein hits were filtered to have at least one peptide ion score of rank 1 above the respective identity threshold (α 0.05). The proteins that did not meet this criterion were rejected. In order to compare the results of both searches the resulting peptide-protein interactions were visualized using Cytoscape (Shannon et al., 2003) as described in Vertommen et al. (2011a) to eliminate false positive results and to reconstruct and annotate the partial sequences of the *Musa* database. In brief, the Excel list of each spot was imported into Cytoscape and a different layout was given to the nodes of peptides and proteins and a different color to the different interactions between the peptides and proteins correlated to the confidence level of identification (ion score). Interactions with an ion score ≥ 40 are displayed in green, <40 in red.

4.3 Results and discussion

4.3.1 Heterotrophic *in vitro* test

Several tests have been performed with different sorbitol concentrations to identify the concentration at which none of the varieties completely stopped growing but they did all show a reduction of their growth (Rukundo et al., 2012). Our main interest lies in identifying varieties that maintain their growth as much as possible even though a mild stress is applied. After a period of 48 days on osmotic stress, the Cachaco variety (ABB) showed the lowest growth reduction. The difference with the AAAh variety Mbwazirume (Kruskal-Wallis one-way analysis of variance by ranks, α 0.05) was significant. While the calculated growth reduction of Cachaco was 63% relative to its control, Mbwazirume displayed a growth reduction of 86%. Popoulou (AAB), Obino L'Ewai (AABp) and Williams (AAA) had intermediate growth reductions of 73%, 71% and 79% respectively (Figure 4.1). The developed screening test with *in vitro* plants has the advantage of being fast and well controlled and is successful at detecting differences in growth reduction. A model will always remain a model and is an attempt to approach reality in an efficient way. Since growth is directly

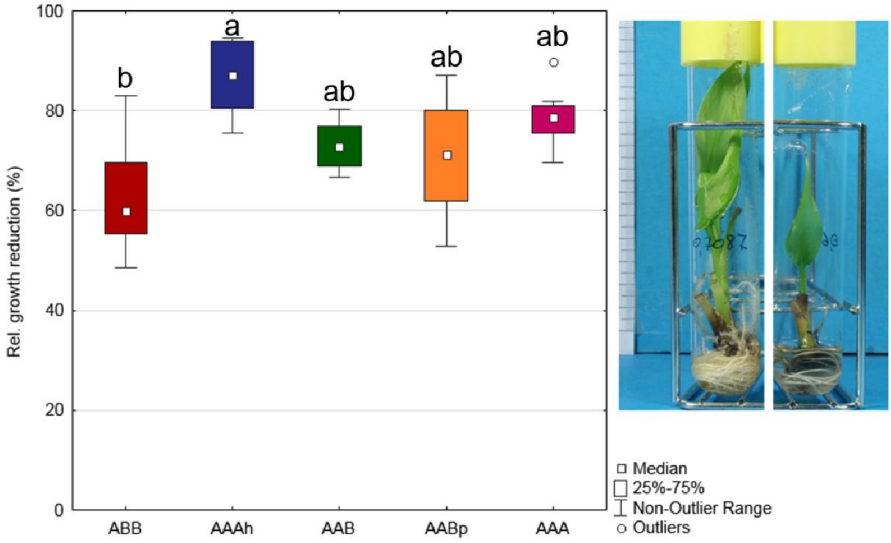

Figure 4.1: Relative growth reduction of sorbitol (0.2 M) stressed plantlets relative to their control after 48 days. Varieties are represented by their genome constitution. Bars marked with the same letter do not differ significantly from each other; a>b (α 0.05, n= 7-8). The picture of the in vitro plantlets shows a control plant on the left and a stressed plant on the right. All roots and leaves have newly been formed during the 48 days.

correlated to yield, growth reduction is an important parameter to judge the stress tolerance of a plant and so the possible yield loss. The ABB variety showed a significantly lower growth reduction than the AAAh variety. Our results are consistent with earlier observations of Rukundo (2009) and confirm that the B genome might be correlated to a higher drought tolerance (Simmonds, 1966; Thomas et al., 1998; Robinson and Sauco, 2010).

4.3.2 Proteomics

As the ABB variety showed the least growth reduction during this osmotic stress, we took a closer look at which proteins were differential between control and stressed plants after 48 days of treatment. This provides an insight at the new equilibrium or homeostasis developed in the stressed plants. After extraction of the leaf proteome and separation of the proteins on gel, 2600 spots were retained in the master gel. A PCA analysis indicates that the control and stressed plants can be discriminated based on the proteome characterization. The most important Principal Component PC1 explains 43.2% of the variation and discriminates the biological samples according to the treatment (Figure 4.2). PC2 is correlated to intra-treatment variability. From the score plot (Figure 4.2), we clearly see that there is more variability in the stressed biological replicates than in the control ones.

Stressed plants can obviously be discriminated based on their proteome, but which proteins are relevant to make the discrimination? To answer this question, a variable importance plot was made based on the loading scores of PC1. Figure 4.3 illustrates that only a few proteins have a very high contribution towards the observed variability between control and stressed samples. Some proteins with a positive PC1 loading (higher abundance during stress conditions) have a high loading score and are important variables. The importance of a variable gradually drops. The same is true for the variables with a negative PC1 loading (higher abundance under control conditions) but we observe that the variables with a positive loading score have a bigger contribution towards the discrimination.

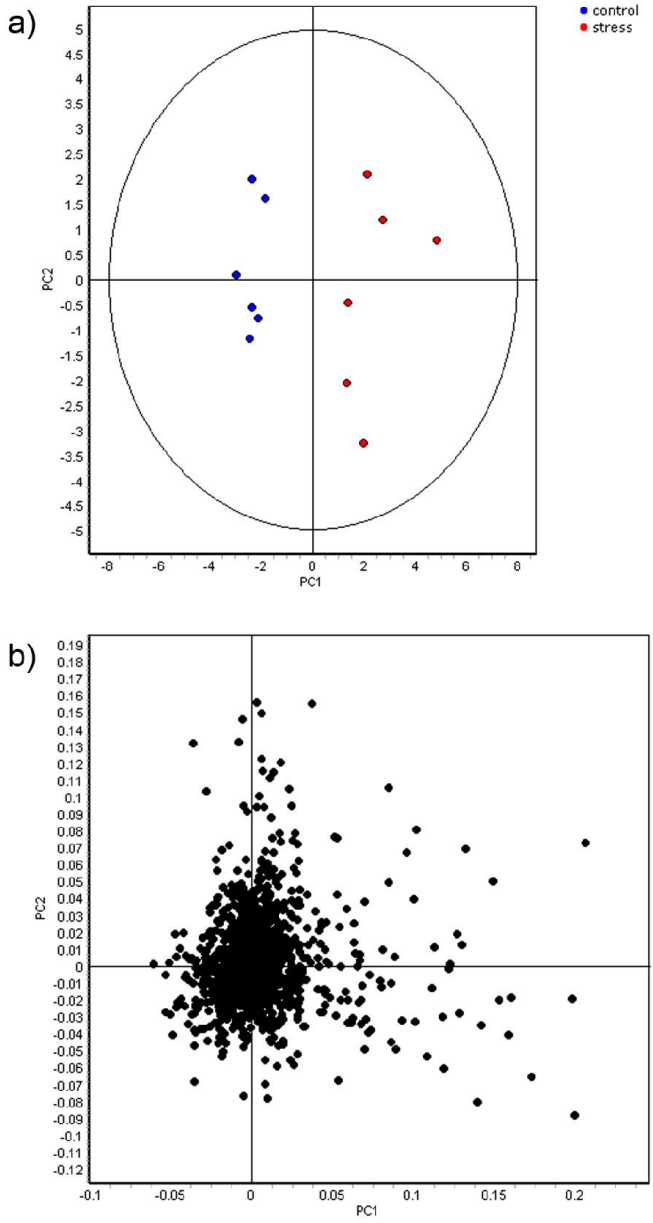

Figure 4.2: PCA score and loading plot. a) PCA score plot with control plants displayed in blue and stressed plants in red. Each dot represents a biological replicate. b) Loading plot. Each dot represents a protein.

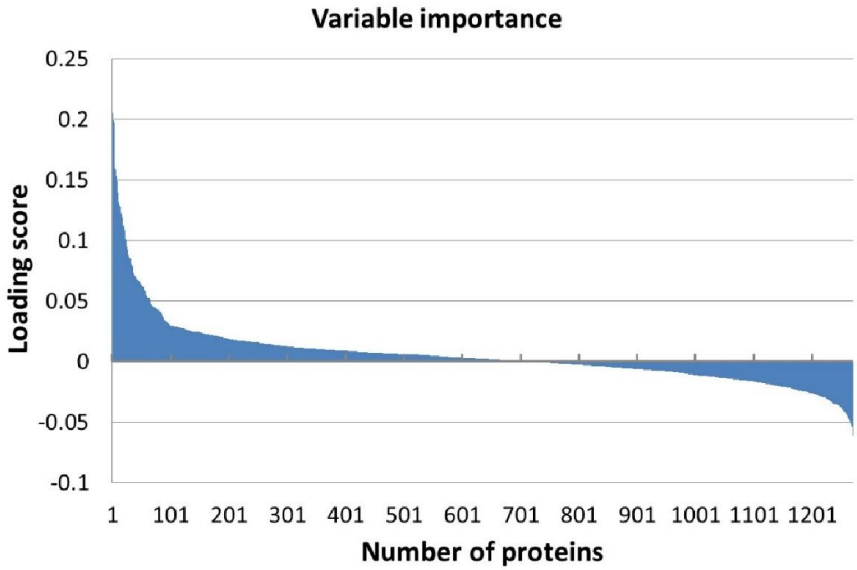

Figure 4.3: Variable importance plot. Variables with a positive loading score for PC1 have a high abundance in stressed samples, variables with a negative loading score for PC1 have a high abundance in control samples.

If we analyze the important variables individually (univariate statistics), we observe that 112 proteins were significantly more abundant in the sorbitol stressed plants and 18 proteins were more significantly abundant in control plants (T-test FDR α 0.05). However, Cy-dyes are very sensitive and some proteins are too low abundant to be efficiently identified. Based on their importance and the abundancy level, 66 differential protein spots were selected for identification and 24 were successfully identified (Table 4.1, Figure 4.4 and Figure 4.5). As in our previous study (Carpentier et al., 2008), we see a correlation to the protein abundance and the success rate of identification (results not shown) although the poor sequencing status of banana also plays a role.

Table 4.1: Overview of the identified differential proteins

Spot ID[§]	Variable importance	PC1	PC2	Protein annotation	n[‡]	T-test	Av. Ratio[•]
62	2	0.199	-0.088	HSP20	12	7.60E-09	12.0
35	3	0.198	-0.019	acidic chitinase	12	9.60E-06	9.5
66	5	0.160	-0.019	PR10	12	7.80E-04	5.8
18	34	0.080	-0.012	isoflavone reductase	12	2.60E-03	2.3
63	38	0.072	-0.039	lectin	12	8.10E-05	2.3
3	48	0.065	0.000	cysteine synthase	10	8.00E-04	2.1
65	73	0.045	-0.023	lectin	12	1.70E-04	1.7
17	84	0.041	-0.014	glutathione S transferase	12	3.90E-04	1.6
39	98	0.031	-0.002	fructose bisphosphate aldolase	12	5.20E-05	1.4
58	140	0.025	-0.012	glutathione reductase	12	6.20E-04	1.4
32	146	0.024	-0.022	isocitrate dehydrogenase[*]	12	2.30E-03	1.4
50	169	0.022	-0.007	fructose bisphosphate aldolase	12	3.10E-04	1.3
21	171	0.022	-0.022	phosphoglucomutase[*]	12	1.50E-03	1.3
31	175	0.022	0.009	glyceraldehyde-3-phosphate dehydrogenase[*]	12	1.30E-03	1.3
2	180	0.021	0.014	transketolase[*]	12	7.70E-03	1.2
10	192	0.020	0.004	unknown protein	10	8.20E-03	1.3
5	209	0.018	-0.007	phosphoglyceromutase[*]	12	9.40E-04	1.2
7	226	0.017	-0.008	S-adenosyl-L-homocysteine hydrolase	12	5.20E-03	1.2
13	9	-0.047	0.006	S-adenosylmethionine synthetase	12	4.10E-06	-1.7
11	16	-0.041	0.006	isocitrate lyase	12	5.60E-04	-1.5
27	22	-0.037	0.005	uroporphyrinogen decarboxylase	12	7.10E-04	-1.5
22	61	-0.028	0.005	eukaryotic initiation factor	12	5.90E-03	-1.3
4	82	-0.025	0.000	eukaryotic initiation factor[*]	10	6.00E-03	-1.3
40	222	-0.014	0.008	methionine synthase[*]	12	8.90E-03	-1.2

[*]Multiple proteins have been identified in this spot. § All spots are displayed in Figure 4.4 and Figure 4.5. ‡ n is the number of gel images of control and stress samples in which the spot was detected. •Average ratio of stress vs control samples

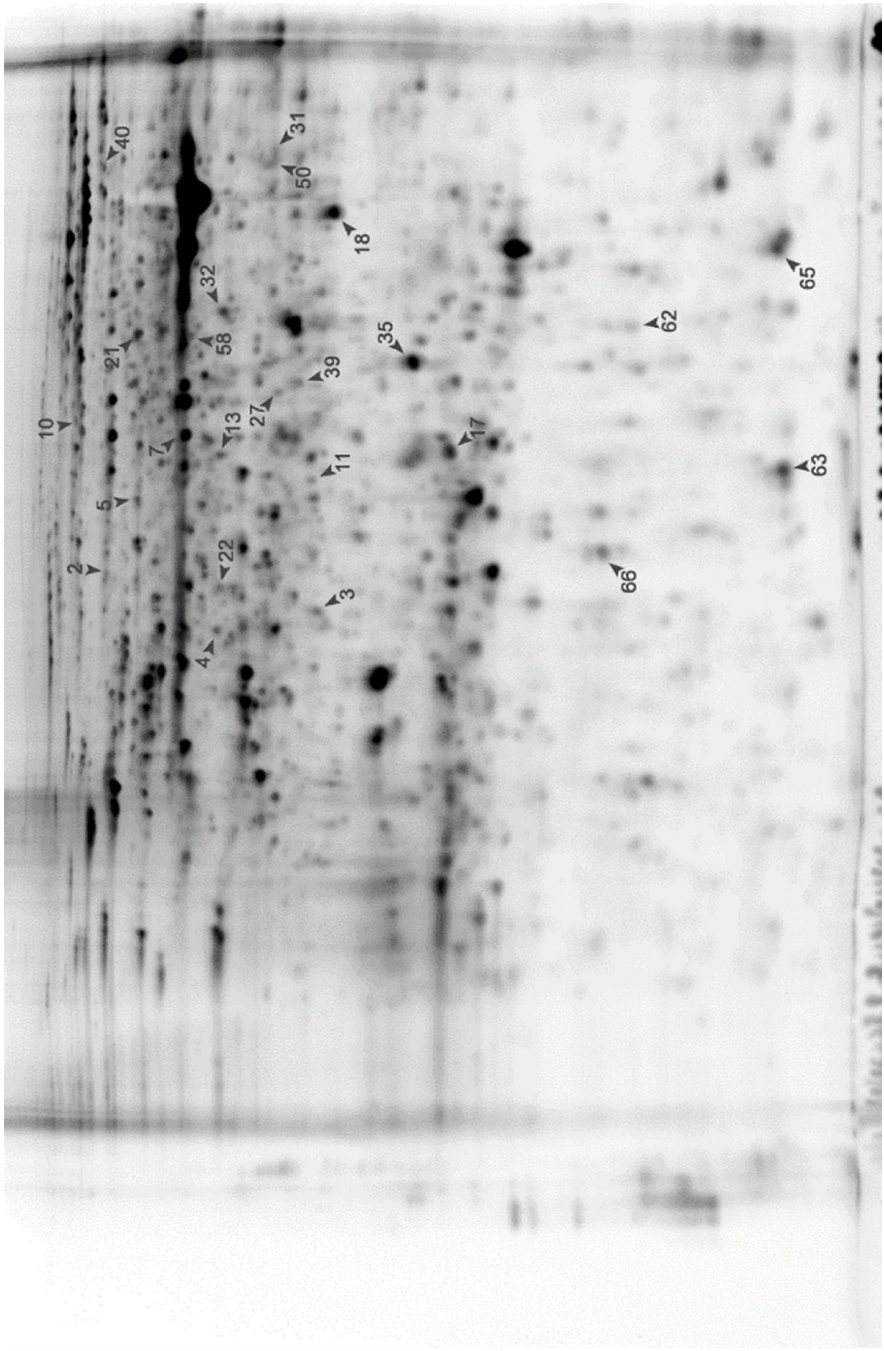

Figure 4.4: Master gel Cy2 labeled (24 cm pI 4-7). Identified proteins are numbered and indicated with an arrowhead (see Table 4.1).

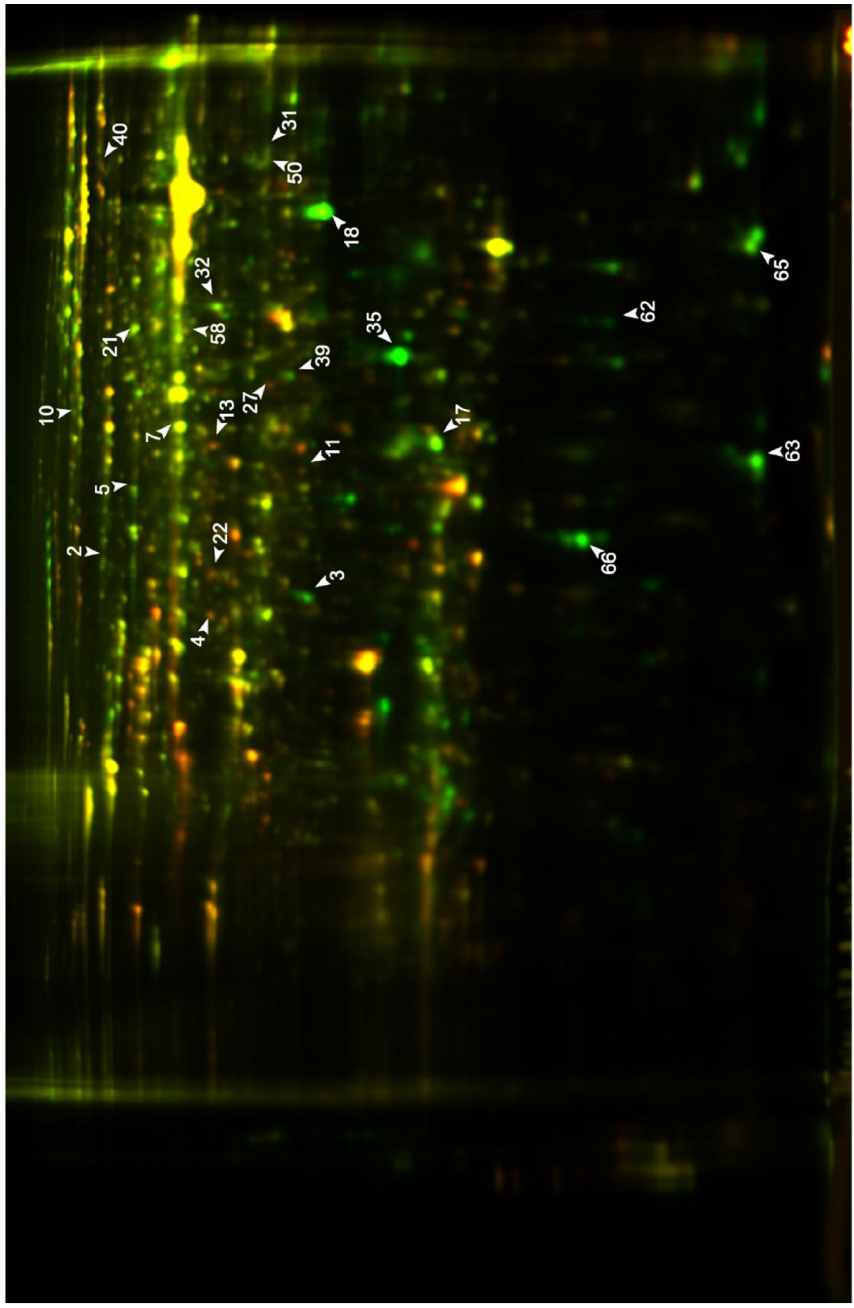

Figure 4.5: Overlay image of Cy3 and Cy5 labels on master gel (24 cm pI 4-7). A stress sample is labeled with Cy3 (green) and a control sample with Cy5 (red). Identified proteins are numbered and indicated with an arrowhead (see Table 4.1).

Despite running the samples on a 3 pI unit strip of 24 cm, some spots still contain multiple proteins. This is due to limitations of the resolution of a 3 pI strip – resolution could still be improved by using zoom strips – and due to the sensitivity of the LC-MS/MS analysis. In automated MALDI-MS/MS analysis often only the 5-10 most abundant peaks are chosen for further fragmentation and lower abundant co-migrating proteins are ignored while here peaks are first separated in time concentrated and analyzed. While in some cases co-migrating proteins create ambiguity, in our cases there is a difference in abundance which can be checked by the number of MS/MS events. Figure 4.6 illustrates a case of 3 possible proteins in one spot. The visualization of the relation between peptides and protein is shown in Cytoscape for spot 32. We clearly see that there are 3 possible proteins: the most abundant protein isocitrate dehydrogenase (gi|3747089 and MusaId000029420), 1-deoxy-D-xylulose 5-phosphate reductoisomerase (isotig05077204791) and class phi glutathione S-transferase (MusaId000019031). Isocitrate dehydrogenase is the most abundant protein since it has not only more peptides but the peptides have also multiple MS/MS events. We report in Table 4.1 the annotation of the most abundant protein assuming that this protein predominates the spot quantification. All peptides/proteins with their corresponding ion/protein scores are listed in Supplementary file 4.1.

We have also observed multiple isoforms of the same protein. Spot 18 contains two different isoforms and very likely a third (Figure 4.7). The peptide STTAPAGQPEK is assigned to MusaId000018332, while the peptide STTAPAGQPEEAK is assigned to MusaId000030279. For both peptides we observed multiple MS/MS events, confirming that indeed both are present. As can be seen in the Cytoscape image, a third cluster is formed with the peptide VVILGDGNTK. Neither MusaId000018332 nor MusaId000030279 give rise to this peptide as they have no lysine before this part of the sequence (Supplementary file 4.2). The tryptic peptide of both these proteins is much larger and exceeds our scan range. The peptide VVILGDGNTK is assigned to MusaId000028304. In contrast to the other 2 proteins, the sequence of MusaId000028304 is only a partial one as the start of the sequence is missing. We will discuss the biological impact of this protein further below.

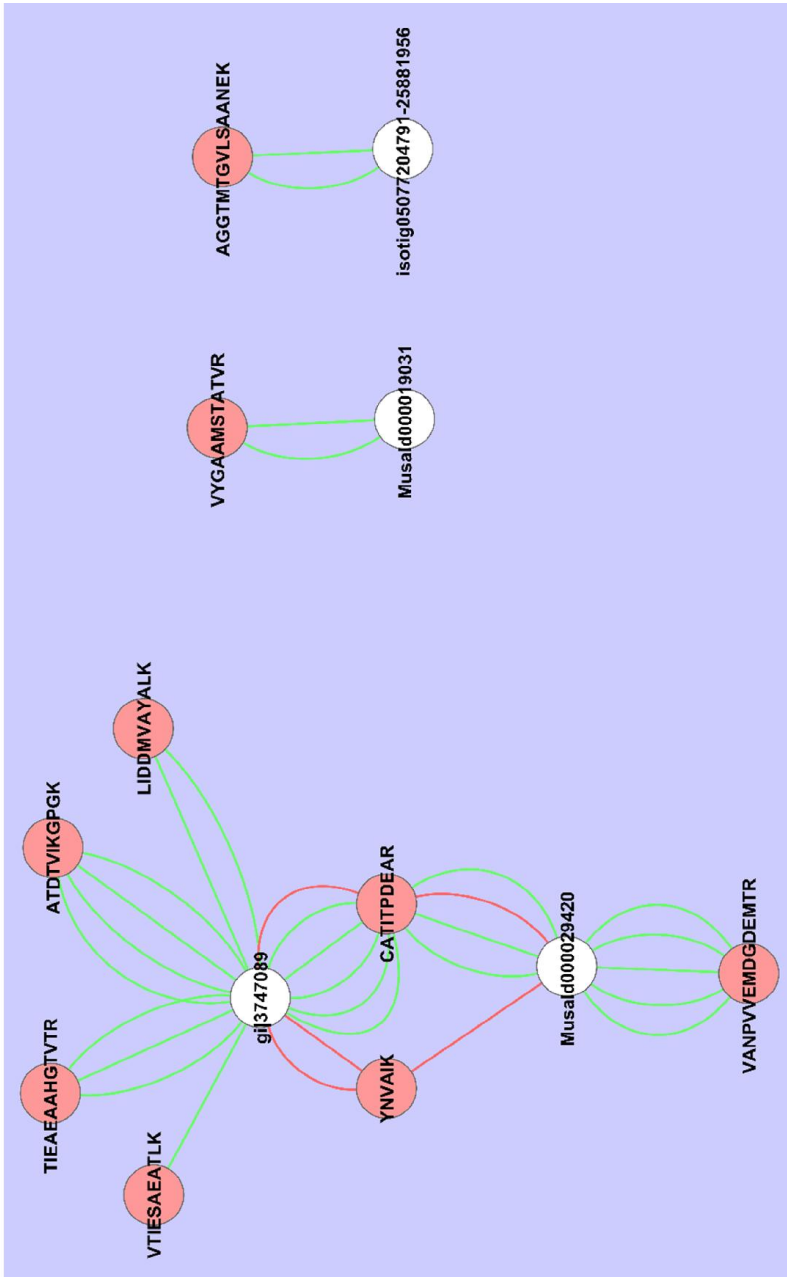

Figure 4.6: Cytoscape representation of spot 32. Redundancy in M/Z values was removed and only the peptide with the highest ion score/proteins score was retained. Interactions with an ion score ≥ 40 are displayed in green, <40 in red. Proteins are represented by white ellipses and peptides by red ellipses

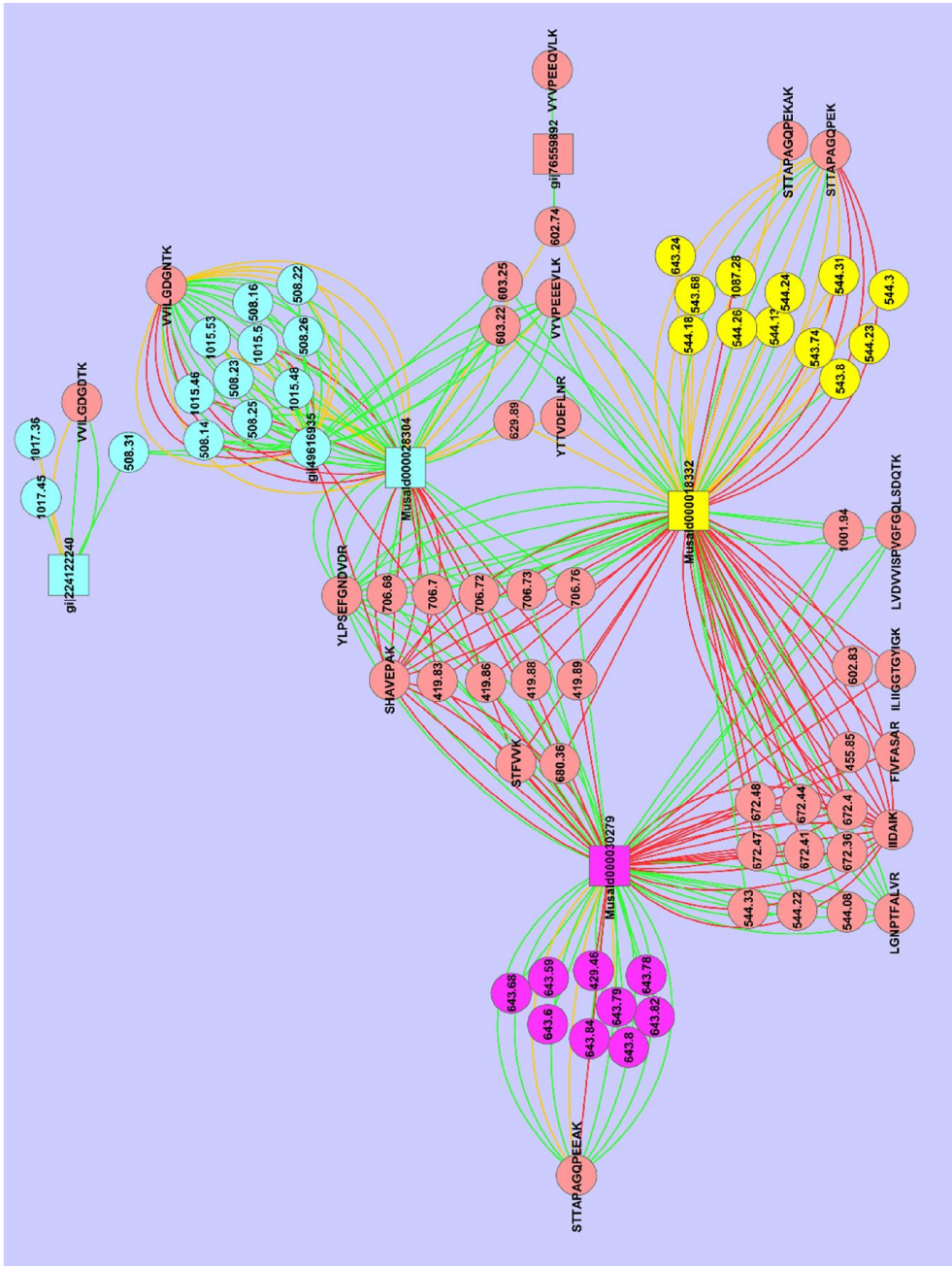

Figure 4.7: Cytoscape representation of peptide-protein interactions of spot 18. Redundancy in M/Z values has not been removed. Interactions with an ion score ≥ 40 are displayed in green, <40 in red. Proteins are represented as squares, peptides and peaks as an ellipse. The three possible isoforms with their tryptic specific peptides are depicted respectively in yellow, purple and green.

4.3.2.1 Proteins involved in stress and reactive oxygen metabolism.

The most important variable that could be identified is a HSP protein of around 20 kDa (spot 62). It contains an alpha crystallin domain (ACD) that is found in small heat shock proteins (sHSPs). sHSPs are molecular chaperones that are generally active as large oligomers consisting of multiple subunits. The *Arabidopsis thaliana* (Ath) AthHsp15.7 is minimally expressed under normal conditions and is strongly induced by heat and oxidative stress. We have calculated that spot 62 is 12 times more abundant in stressed plants. Whether it is here induced directly by osmotic stress or indirectly by oxidative stress remains elusive. We hypothesize it plays an important role in maintaining homeostasis by suppressing protein aggregation.

How can osmotic stress provoke oxidative stress? Tetrapyrroles are natural pigments containing four pyrrole rings and play an important role in the transfer of energy and redox sensing. Chlorophylls are the most abundant tetrapyrroles in plants and are involved in the harvesting of light and its subsequent conversion to chemical energy. Uroporphyrinogen decarboxylase (spot 27) is an enzyme involved in the tetrapyrrole biosynthetic pathway. We observe that this enzyme is less abundant in stressed plants. Reduced levels of uroporphyrinogen decarboxylase slow down the further tetrapyrrole metabolism and increase the level of uroporphyrinogens. Uroporphyrinogens are tetrapyrroles that can be photooxidized, thus triggering photodynamic damage. Mock et al. (1999) characterized the cellular stress responses upon down-regulation of uroporphyrinogen decarboxylase. They observed an accumulation of uroporphyrinogens, increased levels of antioxidant mRNAs and increased activity of enzymes involved in pathogen defense indicating that these cellular reactions upon porphyrinogenesis resemble a hypersensitive reaction after pathogen attack (Mock et al., 1999). We expect that the reduced levels of uroporphyrinogen decarboxylase in stressed plants triggers photodynamic damage and ROS. This might explain why we observed an increased level of typical pathogen defense-related proteins: PR10 (spot 66), lectin (spot 63 and 65), chitinase (spot 35) and proteins involved in reactive oxygen species (ROS) detoxification: isoflavone reductase like protein (spot 18), glutathione reductase (spot 58), cysteine synthase (spot 3), glutathione transferase (spot 17). Enzymes involved in ROS metabolism have been abundantly described in literature. But whether induction of pathogen-related enzymes is a secondary effect of stress (ROS) or whether those enzymes effectively play a role in homeostasis is an interesting question for further research and further annotation of those enzymes. Do those proteins only play a role in pathogen defense or do they have an essential role to play in osmotic tolerance?

We have already mentioned that spot 18 contains two and maybe even three isoforms of the same enzyme. While the *Musa* sequence present in the NCBI database has been annotated as isoflavone reductase, other related reductases,

such as phenylcoumaran benzylic ether reductase also show great similarity. Conserved domain analysis using the conserved domain database confirms the existence of a Rossmann-fold NAD(P)H/NAD(P)(+) binding (NADB) domain. However as to the substrate of the identified reductase we can only speculate. Most likely, like isoflavone reductase, it might play a distinct role in plant antioxidant defense. Isoflavone reductase has been shown to be involved in NAD(P)/NAD(P)H homeostasis (Babiychuk et al., 1995).

4.3.2.2 Proteins involved in energy metabolism and respiration.

Phosphoglucomutase (spot 21), fructose bisphosphate aldolase (spot 39 and 50), glyceraldehyde-3-phosphate dehydrogenase (GAPDH) (spot 31) and phosphoglyceromutase (spot 5) are all part of the glycolysis pathway in plants. Transketolase (spot 2) belongs to the pentose phosphate pathway. The most important function of the glycolysis pathway and the pentose phosphate pathway is to form ATP, reductants (NAD(P)H) and carboskeletons which are building blocks for anabolic pathways. An upregulation of enzymes of this pathway is consistent with our earlier studies on meristems showing that stress creates a higher energy (ATP) and reducing power (NAD(P)H) demand (Carpentier et al., 2007; Carpentier et al., 2010).

The production of reactive oxygen species (ROS), such as O_2^- and H_2O_2, is an unavoidable consequence of normal respiration with the mitochondrial electron transport chain is a major site of ROS production. An enhanced respiration produces higher levels of ROS. The mitochondrial electron transport chain contains two stress upregulated non-proton-pumping NAD(P)H dehydrogenases on each side of the inner membrane which function to limit mitochondrial ROS production (Moller, 2001). Several other enzymes are found in the matrix that, together with small antioxidants such as glutathione, help remove ROS. The antioxidants are kept in a reduced state by matrix NADPH produced by NADP-isocitrate dehydrogenase and the non-proton-pumping transhydrogenase activities.

We have noticed a higher abundance of isocitrate dehydrogenase (spot 32) in stressed plants. Isocitrate lyase (spot 11) is located in the glyoxysome and isocitrate dehydrogenase (spot 32) in the mitochondria. Both enzymes have isocitrate as a substrate and could compete for isocitrate processing. The role of isocitrate lyase has been described especially in oily seeds where the breakdown of fatty acids generates acetyl-CoA. Acetyl-CoA is then used in the glyoxylate cycle, which generates other intermediates that serve as a primary nutrient source prior to the production of sugars from photosynthesis. However, what could be the role of isocitrate lyase in leaf tissue? Compared to our reference control condition, we have noticed that the abundance of isocitrate dehydrogenase is higher and that of

isocitrate lyase lower during stress. This would mean that more isocitrate goes towards respiration than towards fatty acid breakdown. We hypothesize that under normal growing conditions there is plenty of sucrose supplied by the medium that is broken down and stored as fatty acids in a futile cycle. This is not the case during stress conditions and balance of stressed plants is more towards respiration to maintain homeostasis.

Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) has been described in many stress studies (Kosova et al., 2011). GAPDH generates NADH from NAD+. Overexpression of GAPDHa in Arabidopsis protoplasts strongly suppressed heat shock-induced H_2O_2 production and cell death (Baek et al., 2008).

4.4 Conclusions

We conclude that an *in vitro* growth model is useful to screen the *Musa* biodiversity for tolerant varieties. The interesting varieties (including sensitive genotypes) will be further investigated and validated under less artificial conditions to study drought tolerance mechanisms. Proteomics is successful in getting an insight into the homeostasis. The proteome analysis clearly shows that there is a new balance in the stressed plants and that the respiration, metabolism of ROS and several dehydrogenases involved in NAD/NADH homeostasis play an important role. This research is a first important step in the understanding of homeostasis and brings new key questions. In the future, we need to elucidate the role of the different isoforms and of poorly annotated *Musa* specific proteins of multiple genotypes and need to clarify the up-regulation of at first sight pathogen-related proteins. A dynamic stress study of different genotypes combined with supervised multivariate analysis needs to clarify which genotypic differences contribute to stress tolerance.

Chapter 5

Characterization of the HSP70 family during osmotic stress

YL and SC performed the proteomics of the meristem experiments. A-CV performed the proteomics of the plant roots and analyzed the proteomics data. Mass spectrometry was performed by SYBIOMA. A-CV curated the *Musa* HSP70 sequences, analyzed the mass spectrometry data and performed the ubiquitination site analysis and promotor analysis of the HSP70 sequences.

5.1 Introduction

Plant HSP70s (70 kilodalton heat shock proteins) play important roles in protein folding, protein import and translocation processes. They are present in the cytosol as well as in mitochondria, chloroplasts and the endoplasmic reticulum (Boston et al., 1996; Miernyk, 1997; Wang et al., 2004). HSP70s associated with the endoplasmic reticulum are usually called luminal binding HSP70s or BiPs. The cytoplasmic HSP70s can be either constitutively expressed or only when the plant is stressed. The constitutively expressed HSP70s are also known as the cognate HSP70s or HSC70s. The exact function of the different HSP70 isoforms is most likely determined by the location in the cell and interaction with co-chaperones. While cytoplasmic HSP70s are involved in the folding of *de novo* synthesized proteins and maintaining precursor proteins in the correct state, mitochondrial, chloroplastic and luminal HSP70s are involved in the precursor protein import and translocation and folding in the respective organelle (Miernyk, 1997; Wang et al., 2004). HSP70s are, as their name implies, most known for their response to heat shock but several HSP70s also react to cold stress, salt stress, drought stress, light and even biotic stresses. The model plant *Arabidopsis* has 14 HSP70s and the most stress responsive one at the transcriptome level is AtHSP70-4 which is correlated to several abiotic and biotic treatments (Swindell et al., 2007). Several HSP70s in rice react differently to drought, salt, heat and light as their mRNA is sometimes highly up regulated by certain stresses and unaffected or even down regulated by others (Jung et al., 2013). In a transcriptomics study in spinach a BiP (luminal HSP70) was upregulated in response to cold (Anderson et al., 1994). In contrast to the differential regulation of the mRNA, these authors discovered that the protein level remained constant, demonstrating the importance of also studying the HSP70s at the protein level and not just at the transcriptome level. Most plant research has focused on *Arabidopsis thaliana* or plant species that have already been characterized to a great extent, the 'new models', such as rice and soybean. But many crops which are essential for food are complex due to their (allo)polyploid heterogeneous genome. Often, they possess many different characteristics that are unique to their species and cannot always be approached via a model plant.

2DE is still the most preferred way to characterize the proteome of a non-model species, but 2D gels often contain multiple spots that are functionally annotated identically. Without a sequenced genome, it is almost impossible to go beyond this general annotation. Gene families (paralogs), allelic variations and/or posttranslational modifications (PTMs) are at the origin of different pI protein species. HSP70s were already identified earlier in proteomics studies by our research group and others as an important player during stress but insight into the protein

polymorphisms, caused by paralogs, allelic variants and/or PTMs, remained unknown (Taylor et al., 2005; Carpentier et al., 2007; Vincent et al., 2007; Carpentier et al., 2010; Kjellsen et al., 2010; Abreu et al., 2013). In the framework of cryopreserving the *Musa* collection several osmotic acclimation studies were performed in the past on meristem tissue cultures. Several HSP70 spots with differential responses were identified in those proteomics studies, but going beyond the level of identifying them as belonging to HSP70 family was impossible using cross-species identification and/or EST databases (Carpentier et al., 2007; Carpentier et al., 2010). With the recent publications of the reference A genome and a draft B genome (D'Hont et al., 2012; Davey et al., 2013), we went beyond the level of simply identifying the trail of spots as belonging to the HSP70 gene family in an ABB variety and annotated the different paralogs and allelic variants. We performed a new experiment on meristems and also characterized the behavior of the HSP70 spots on the plant level (roots) after 0, 1, 4 and 14 days of osmotic stress using 2D-DIGE. We confirmed that HSP70s play a role in osmotic stress and we identified one particular spot that specifically reacted towards the osmotic stress in both meristems and in roots. To understand what was special about that particular spot, we characterized the spots via 2DE LC-MS/MS and investigated the different isoforms. To our knowledge this is the first time a proteomics approach has led to the exploration of a protein family at the paralog and allelic level in a crop. Furthermore, we identified a specific osmotic responsive HSP70 protein species. To gain an insight into this differential expression, we also performed a promoter analysis of all the identified isoforms.

5.2 Experimental procedures

5.2.1 Analysis of the *Musa* HSP70 family

Musa HSP70 nucleotide and protein sequences were obtained from GreenPhyl and the Banana Genome Hub (D'Hont et al., 2012; Droc et al., 2013). Since many HSP70 genes were incorrectly predicted, all HSP70 A genome sequences were manually curated as well as the cytoplasmic and luminal HSP70 B genome sequences. Cytoplasmic HSP70 from rice and *Arabidopsis thaliana* were retrieved from Greenphyl with the accessions as described by Jung et al. (Jung et al., 2013). Alignments of protein sequences were created using the ClustalX 2.1 software (Larkin et al., 2007). Phylogenetic trees were constructed using the same software using the neighbor-joining algorithm with 1,000 replicate bootstrap tests. Trees were visualized with njplot (Perrière and Gouy, 1996).

For the prediction of the potential protein ubiquitination sites, all cytoplasmic protein sequences encoded by the A genome were submitted to UbiPred and UbPred (Tung and Ho, 2008; Radivojac et al., 2010). For the promoter analysis the sequence between the start codon ATG of the HSP70 genes and the sequence of the previous gene was analyzed up to a maximum of 3kb from the start codon. To find the *cis*-regulatory elements that possibly influence the expression, PlantCARE and PLACE softwares were used (Higo et al., 1999; Lescot et al., 2002).

5.2.2 *In vitro* meristem stress tests

In vitro plants of the selected variety Cachaco (ABB, ITC 0643) were supplied by the International Transit Centre of Bioversity International. Multiple shoot meristem cultures were initiated as described by Strosse et al. (Strosse et al., 2006) and maintained on the standard control medium (MS medium supplemented with benzylaminopurine). All cultures were kept in the dark at 25-27 °C. A stress test was started by adding 0.31 M sucrose to the standard medium. Tissue samples of stressed meristem cultures were taken and frozen after 0, 1, 4 and 14 days. All samples were stored at -80 °C.

5.2.3 Plant root stress test

In vitro plants of the selected variety Cachaco (ABB, ITC 0643) were supplied by the International Transit Centre of Bioversity International. The plants were grown in a phytotron (Sanyo, MLR-351H). The humidity and temperature were kept constant at 75 % and 25 °C respectively. A 12 h/12 h light/dark period with an average light intensity of 183 ± 29 µmol photons m^{-2} s^{-1} was maintained throughout the experiment. After five weeks an osmotic stress test was started by adding 0.21 M sorbitol to the MSR medium (MSR medium according to Voets et al. (2005)). Root samples of control plants were taken and frozen in liquid nitrogen at the start of the experiment and after 4 days. Root samples of sorbitol stressed plants were taken and frozen after 0, 1, 4 and 14 days. All samples were stored at -80 °C.

5.2.4 Proteomics

Meristem and root proteins were extracted and analyzed using the phenol extraction/ammonium acetate precipitation protocol reported by Carpentier et al. (2005). Fifty µg of proteins were labeled with Cy2, Cy3 and Cy5 (GE Healthcare) for a total of 150 µg protein per gel, separated on gel and scanned according to Carpentier et al. (Carpentier et al., 2009). Data were analyzed using the DeCyder software version 7.0 (GE Healthcare). Statistical analysis of the standardized abundance of spots was performed in DeCyder. Statistical analysis of the raw spot intensities was

performed using ANOVA in STATISTICA software 10 on the log of the peak height of the internal standard samples exported from DeCyder. For protein identification, gel pieces were extracted based on the protocol of Shevchenko et al. (2006) for the in-gel reduction, alkylation and destaining of the proteins. The destaining step was performed twice after which the gel pieces were covered with 3 µL of 0.1 µg/µL trypsin and 47 µL trypsin buffer (25 µM ammonium carbonate, 10% acetonitrile (ACN)). Digestion was performed overnight at 37°C. Peptides were extracted by adding 100 µL 5 % ACN in 0.1 % FA, vortexing, centrifuging and sonicating for 5 min after which the supernatant is removed to a new eppendorf tube. The whole peptide extraction process is repeated twice with 50 µL 10 % ACN in 0.1 % FA the first time and 50 µL 95 % ACN and 5 % FA the last time. The accumulated supernatant was then dried in a vacuum centrifuge and stored at -20 °C. Before analysis, the samples were resuspended in 0.1 % FA and 5 % ACN, desalted using C18 Zip Tips (Millipore) and eluted in 10 µl Milli-Q water with 0.1 % FA and 60 % ACN, dried in a vacuum centrifuge and resuspended in 0.1 % FA and 5 % ACN.

The HPLC-MS/MS analysis was performed on a Q Exactive Orbitrap mass spectrometer (Thermo Scientific, USA). The samples (5 µL) were injected and separated on an Ultimate 3000 HPLC system (Dionex, Thermo Scientific) equipped with a C18 PepMap100 precolumn (5 µm, 300 µm x 5 mm, Thermo Scientific) and an EasySpray C18 column (3 µm, 75 µm x 15 cm, Thermo Scientific) using a gradient of 5 % to 20 % ACN in 0.1 % FA in 10 min followed by a gradient of 10 % to 35 % ACN in 0.1% FA in 4 minutes and then a final gradient from 35 % to 95 % ACN in 0.1 % FA in 2.5 min. The flow-rate was set at 250 µL/min. The Q Exactive was operated in positive ion mode with a nanospray voltage of 1.5 kV and a source temperature of 250 °C. ProteoMAss LTQ/FT-Hybrid ESI Pos. Mode CalMix (MSCAL5-1EA SUPELCO, Sigma-Aldrich) was used as an external calibrant and the lock mass 445.12003 as an internal calibrant. The instrument was operated in data-dependent acquisition (DDA) mode with a survey MS scan at a resolution of 70,000 (fwhm at m/z 200) for the mass range of m/z 350-1800 for precursor ions, followed by MS/MS scans of the top 10 most intense peaks with +2, +3 and +4 charged ions above a threshold ion count of 16,000 at 35,000 resolution using a normalized collision energy of 28 eV with an isolation window of 3.0 m/z and dynamic exclusion of 10 s. All data were acquired with Xcalibur 2.2 software (Thermo Scientific). For identification, all raw data were converted into mgf files using Progenesis v4.1 (Nonlinear Dynamics, UK). The spectra were searched using Mascot (version 2.2.04) against our in-house *Musa* database (76,220 sequences) containing all the protein sequences of the published A and B genome plus contaminant sequences (trypsin and keratin). Redundancy was eliminated from the database using the program cdhit (Li et al., 2001). If both A and B isoforms were identical, the B genome isoform was eliminated. The original HSP70

protein sequences were removed and replaced by the manually curated HSP70 sequences. Search parameters were set at: tryptic digestion, one miscleavage allowed, 10 ppm precursor mass tolerance and 0.02 Da for fragment ion tolerance with a fixed modification of cysteine carbamidomethylation and a variable modification of methionine oxidation.

An isoform was retained as positively identified in a spot if at least one tryptic specific peptide was found with an ion score higher than the Mascot identification score. Cytoscape v3.0 software was used to visualize tryptic specific peptides (Shannon et al., 2003; Vertommen et al., 2011a; Carpentier and America, 2014). Two identified paralogs have allelic variants that cannot be discerned from each other in this analysis. The chromosome 7 allelic variants, 7T15160 and 7_G19958, are identical at the protein level and are reported uniformly as 7T15160 since the 7_G19958 was removed from the database. For the chromosome 5 and 6 allelic variants as well as the luminal chromosome 9 allelic variants, only peptides were observed that are shared between the two allelic variants and the proteotypic peptide is not allelic variant specific. Therefore our analysis for the cytoplasmic chromosome 5, 6 and 7 isoforms and luminal chromosome 9 isoforms remains at the paralog level. To quantify the different protein species in each spot, the Mascot emPAI was exported and the ion intensity of the proteotypic peptide for each peptide was analyzed in Progenesis v4.1. Moreover, for all isoforms positively identified in at least one spot, we searched the unidentified MS/MS spectra in each spot in which they were not identified by performing a manual SRM approach. The ion intensity for a MS/MS spectrum was added to the quantification when the peptide fragment mass corresponded to the proteotypic peptide and a specific signature m/z was identified in the MS/MS spectrum.

5.3 Results

5.3.1 Overview of HSP70 family

GreenPhyl (Rouard et al., 2011), a database for comparative and functional genomics in plants, predicts that the banana reference A genome contains 47 genes in the HSP70 superfamily. After initial analysis, it was clear that the amino acid sequences of most HSP70s were not correctly predicted. Out of the 47 sequences suggested by GreenPhyl, a total of 10 suggested sequences showed the most resemblance to the HSP110/SSE subfamily of the HSP70 superfamily and these were not further analyzed in the scope of this study. Four sequences were functionally misannotated and 11 sequences were pseudogenes of HSP70 genes and not full sequences. Manual curation of the structural annotation consisted mainly of removing inexistent introns

and merging wrongfully separated accessions. After manual curation of the sequences, we identified several additional peptides through proteomics which proved the accuracy of this manual correction of the genome prediction (Table 5.1). Indeed, proteomics can be used to identify genes and splice variants, validate predicted exons and genes as well as to study genome variation (Renuse et al., 2011). As only the 8T20830 sequence was correct before manual curation, 8T20830 seemed to be the prevalent isoform in most spots. This analysis was corrected after the curation (see 5.3.2).

Table 5.1 Proteogenomics: Peptides identified after curation of the A genome and the B HSP70 sequences which were not present in any of the uncurated sequences.

Isoform ID	Peptide identified after curation	m/z	Best ion score
2T16250	MYQGAGGGMGGGMDEDIPSTGGSSGAGPK	893.032	41.06
6T34210	NALENYAYNMR	687.809	51.82
	KIEDAIEK	473.267	44.45
	ELEGICNPIIAK	678.864	42.76
	MYQGAGADMAGGMDEDGPTTGGSSAGPK	888.689	105.68
7T15160	TTPSYVAFTDSER	737.347	86.59
	NTINDDKIASK	406.882	57.13
	NALENYAYNMR	687.809	55.16
9T03960	NALENYAYNMR	687.810	51.82
10T00900	NALENYAYNMR	687.809	55.16
9T23710	NQLETYVYNMK	709.835	45.92
8_G23681	NQVAMNPINTVFDAK	839.417	107.23
9_G27669	DAVVTVPAYFNDAQR	833.415	18.6

After curation of the 47 predicted sequences, a total of 20 *Musa* HSP70 sequences was retained. HSP70 sequences are present on all chromosomes of the *Musa* genome, except chromosome 1 and 11, with up to five sequences on chromosome 6 (Figure 5.1).

Four clusters of isoforms can be distinguished within the phylogenetic HSP70 tree which correspond to their respective locations in four compartments of the plant cell (Figure 5.2). This is supported by the presence of the specific C-terminal motif in almost all of these sequences. Cluster 3 contains all the cytoplasmic HSP70s with the cytoplasmic motif (EEVD) being conserved in 9 out of 11 sequences (Guy and Li, 1998; Sung et al., 2001a). The luminal motif (HDEL) is completely conserved in all sequences within cluster 4. The chloroplastic and mitochondrial motifs (respectively PEGDVIDADFTDSK and PEAEYEEAKK) are not completely but mostly conserved in all the sequences in cluster 2 and cluster 1 respectively.

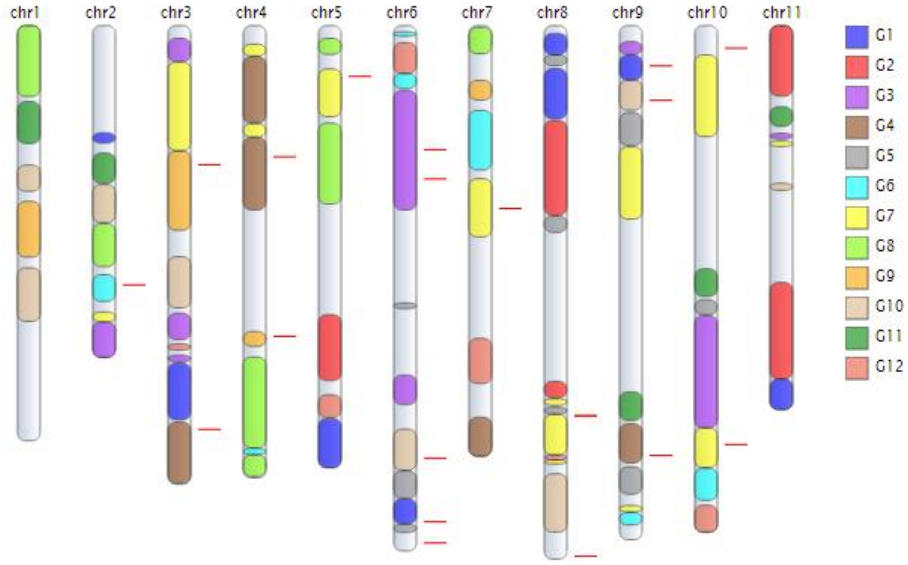

Figure 5.1: Karyotype representation of Musa from the Banana Genome Hub. Locations of the HSP70 Musa genes are represented by red bars. The 12 Musa ancestral blocks (G1-12) are represented by the colored boxes within the chromosomes. Duplicated gene clusters were tentatively assembled into these 12 ancestral blocks and represent the Musa genome before the last two whole genome duplications. On chromosome 8 two genes (GSMUA_Achr8T20830_001 and GSMUA_Achr8T20840) are represented by a single (thick) bar as they are located too close together to be discernable.

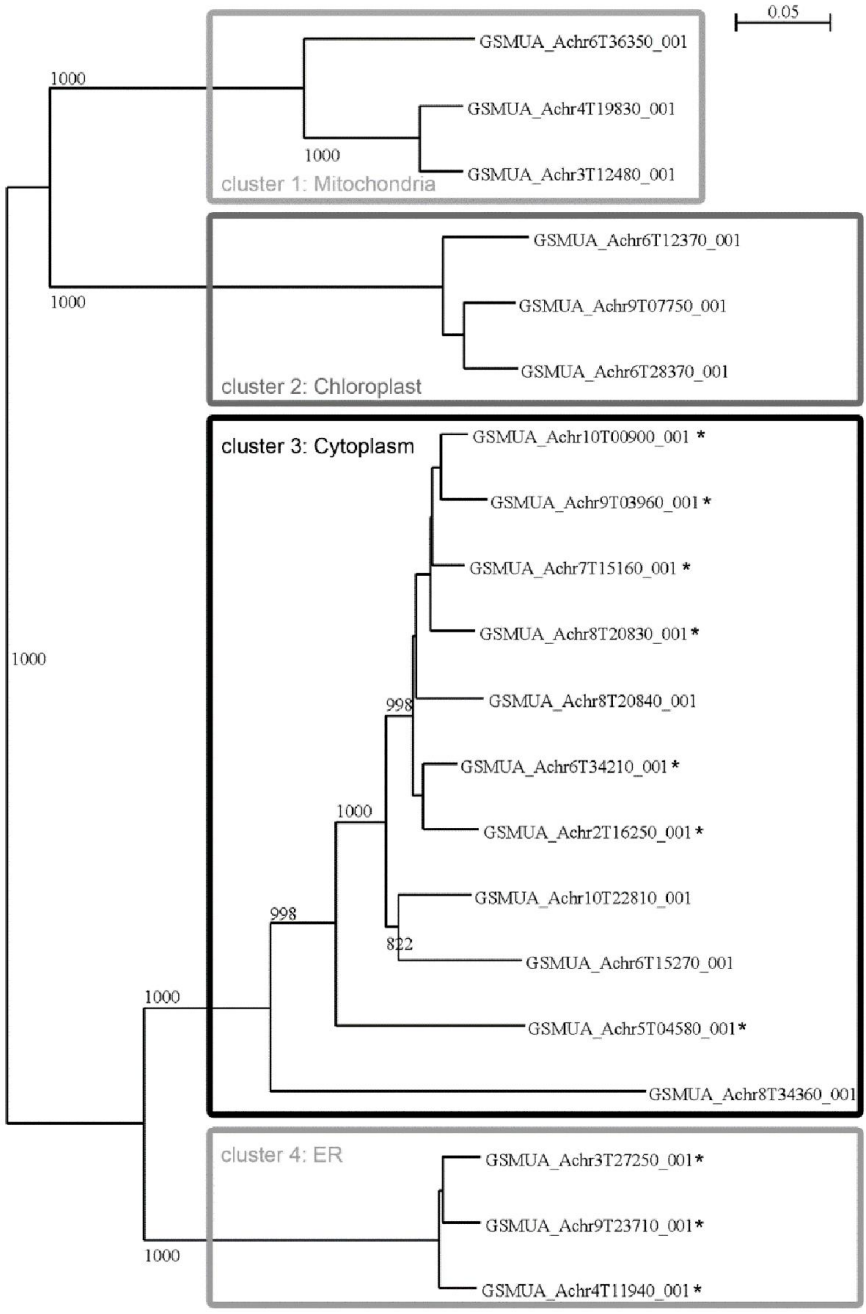

*Figure 5.2: Phylogenetic relationship between all curated HSP70 protein sequences encoded by the Musa A genome. Sequences were aligned using ClustalX and a neighbor-joining tree was constructed with 1000 replicate bootstrap. The bootstrap values between major groups are indicated. Gray boxes indicate the four major sub-clusters corresponding to the HSP70 localization in the cell. Identified paralogs in the proteomic analysis are indicated by *.*

5.3.2 Proteomics

5.3.2.1 Meristems

A trail of 6 spots was identified as HSP70-like proteins in two experiments performed in the framework of the cryopreservation research (Carpentier et al., 2007; Carpentier et al., 2010). In 2010, one HSP70 spot ('spot 3001') was significantly more abundant at 4 days under both sucrose and sorbitol stress. The spots in the trail were all identified as belonging to the HSP70 family but identifications at the isoform level were at that time not possible using cross-species and EST databases. After the publication of the reference genome, we repeated the experiment once more and also included an analysis after 1 day of stress resulting in a 0, 1, 4 and 14 kinetic analysis.

The same trail of HSP70-like spots (1, 2, 3, 4, 5 and 6) was detected (Figure 5.3). Spot 2, based on pattern matching the equivalent of 'spot 2710', was the most intense spot of the six ($\alpha<0.01$) (Figure 5.4) and spot 5 (corresponding to 'spot 3001' in the 2010 study) was again more abundant under sucrose stress ($\alpha<0.01$) (Figure 5.5). A three replicate control versus stress analysis at 4 days confirmed yet again that the abundance of spot 5 significantly increased during osmotic stress ($\alpha<0.01$). A variety with a different genome constitution, Mbwazirume (AAA) was included in the same experiment and spot 5 was similarly more abundant in stress conditions compared to control conditions. (Figure 5.6).

5.3.2.2 Roots

The meristem acclimation experiments were performed in the framework of cryopreserving the *Musa* biodiversity. To investigate whether HSP70s also play a role in the acclimation of roots to osmotic stress a dynamic stress experiment was performed on autotrophic plants.

The analysis of the roots revealed that the abundance of spot 5 similarly increased in time when the roots were subjected to osmotic stress (Figure 5.7). The abundance of other spots showed a similar profile as in the meristems. A six replicate control versus stress analysis at 4 days confirmed once more that the abundance of spot 5 was higher than the control treatment ($\alpha<0.1$) (Figure 5.8).

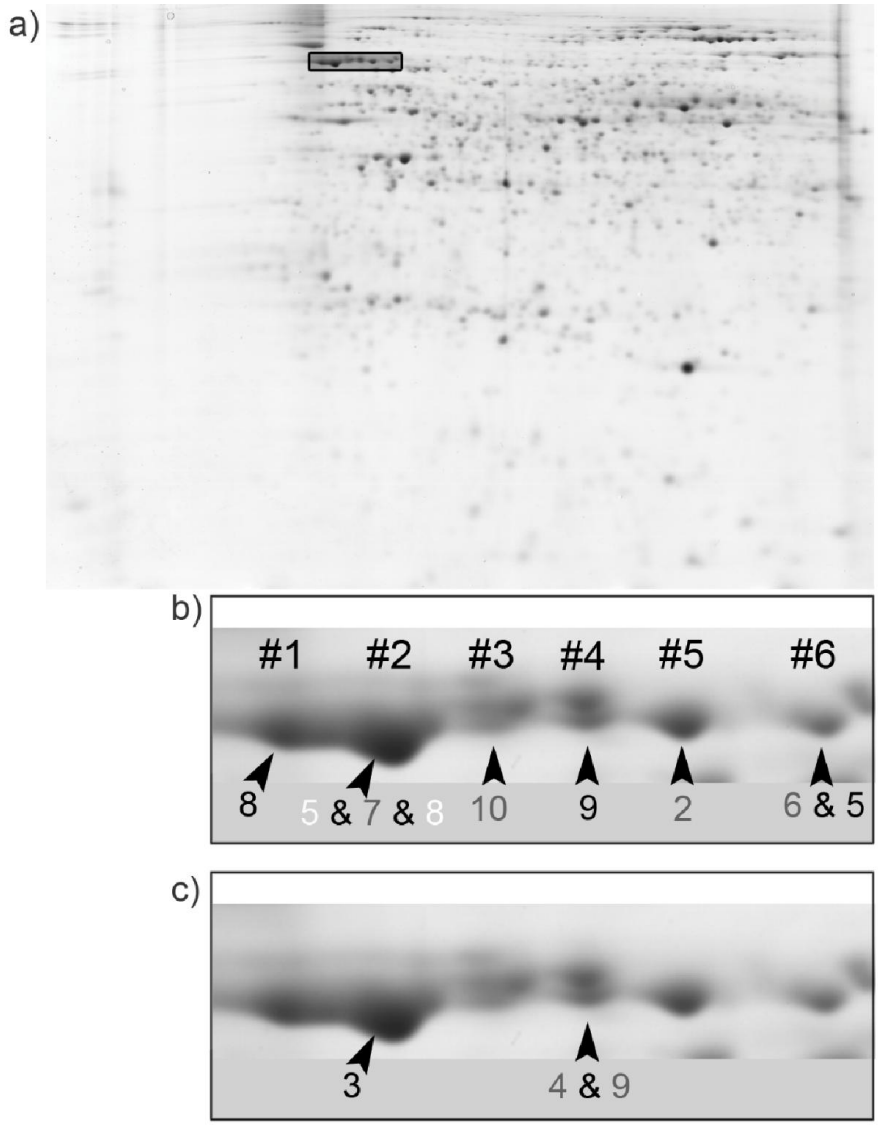

Figure 5.3: a) Representative gel (pI 4-7, 24 cm) with relevant outtake. b) Cytoplasmic isoforms, represented by their chromosome number as in Table 5.2, are indicated with an arrow at the spot where the maximum intensity of their proteotypic peptide is located. A genome allelic variants are represented in white, B genome allelic variants in black and co-localized or indiscernible A and B allelic variants in grey. Spot numbers are indicated with # c) Luminal isoforms, represented by their chromosome number as in Table 5.2, are indicated with an arrow at the spot where the maximum intensity of their proteotypic peptide is located. Allelic variants are represented as in b.

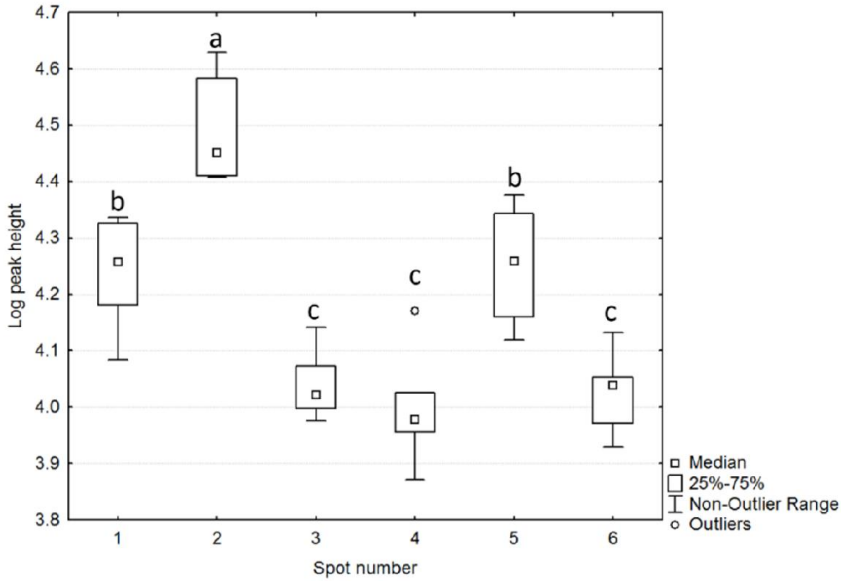

Figure 5.4: Analysis of intensity of the different spots from the meristem experiment using log peak height of the internal standard pools (α<0.01, n=6, a>b>c, outlier = 1.5x interquartile length).

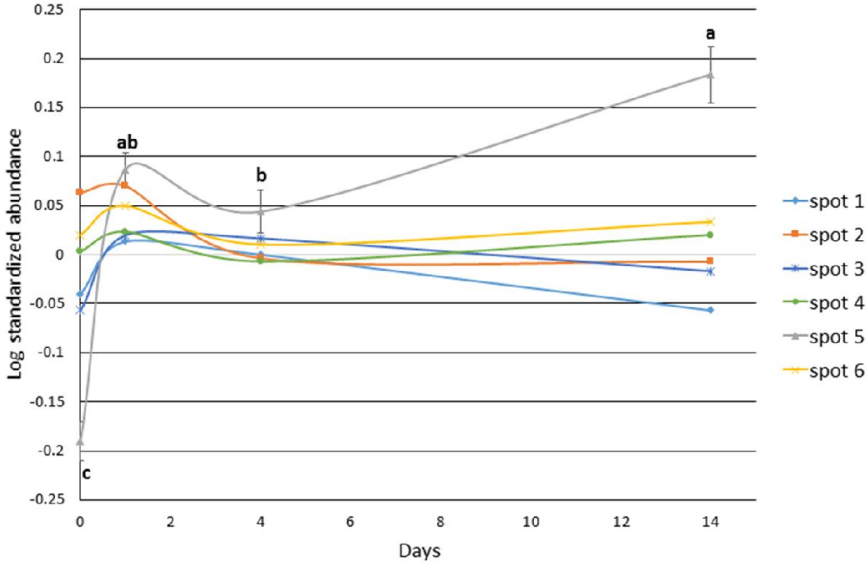

Figure 5.5: Dynamic abundance profiles of the meristem HSP70 spots 1, 2, 3, 4, 5 and 6 after 0, 1, 4 and 14 days of stress (n=3). For spot 1, 2, 3, 4 and 6 mean values are indicated. For spot 5 mean values ± SE are represented. Sample points with the same letter do not differ significantly from each other (α<0.01, a>b>c).

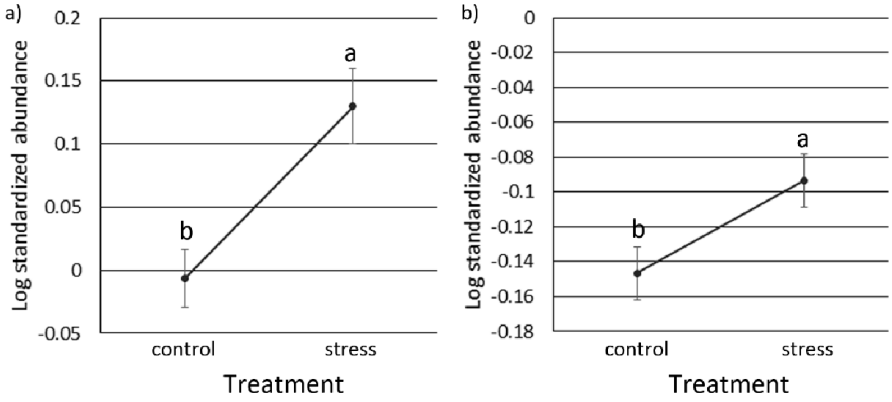

Figure 5.6: Abundance profile of the meristem HSP70 spot 5 after 4 days of control and stress treatment (α<0.01, a>b, n=3). Mean values ± SE are represented. a) Abundance profile of Cachaco (ABB) meristems. b) Abundance profile of Mbwazirume (AAA) meristems.

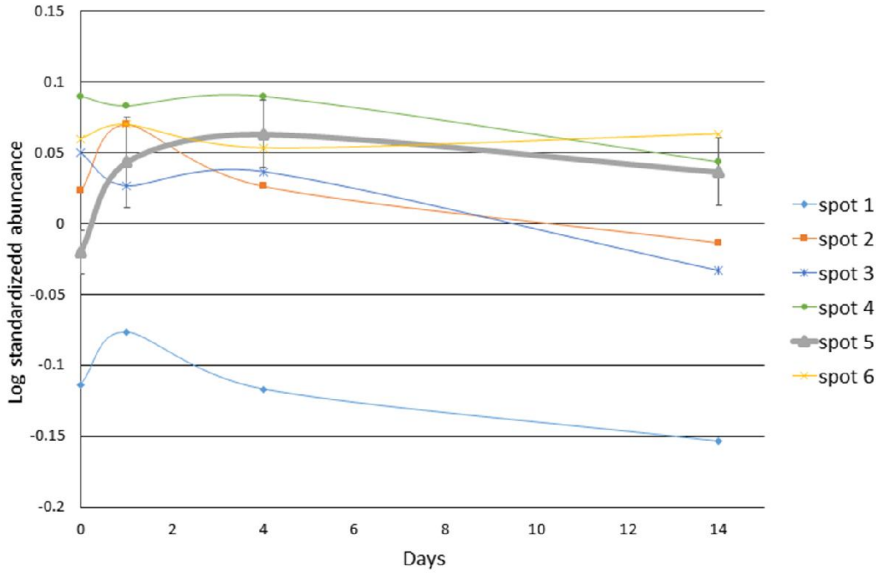

Figure 5.7: Dynamic abundance profiles of the root HSP70 spots 1, 2, 3, 4, 5 and 6 after 0, 1, 4 and 14 days of stress (n=3). For spot 1, 2, 3, 4 and 6 mean values are indicated. For spot 5 mean values ± SE are represented.

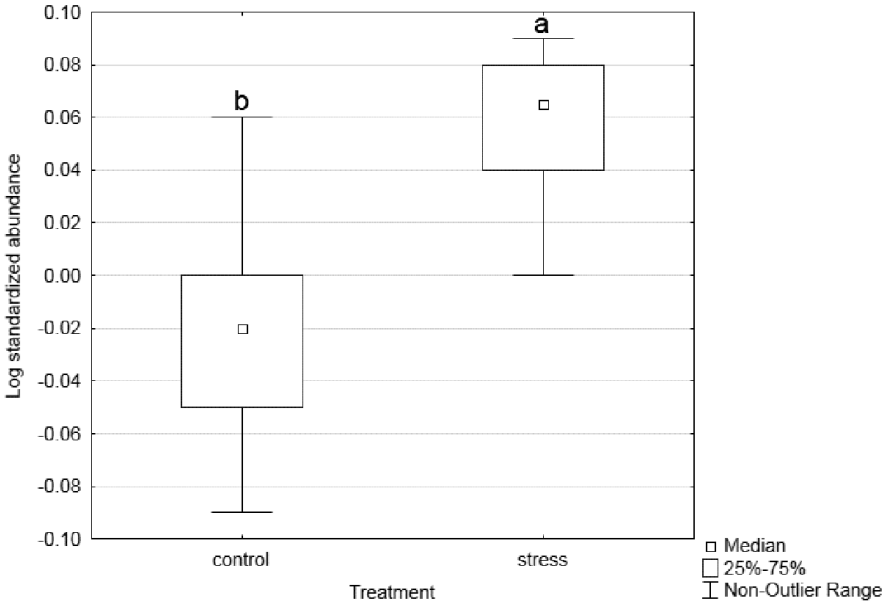

Figure 5.8: Abundance profile of the root HSP70 spot 5 after 4 days of control and stress treatment (α<0.1, a>b, n=6).

5.3.2.3 LC-MS/MS

Blind clustering of parent masses from a first MALDI-MS/MS analysis with SPECLUST (Alm et al., 2006) showed 3 big clusters: spot 1 and 2 (dominated by acidic cytoplasmic isoforms), spot 3 and 4 (containing increased amounts of luminal isoforms) and spot 5 and 6 (dominated by basic cytoplasmic isoforms) (Figure 5.9).

Since MALDI MS-based analysis was not sufficient to detect proteotypic peptides (results not shown), the digested protein mixture was further analyzed via LC-MS/MS to get an insight into the composition of the different spots, to understand why one spot (spot 2) is more abundant than the others and why one spot (spot 5) is more responsive to osmotic stress. Different protein isoforms expressed from the same genome are called paralogs as they may have arisen from gene or segmental duplication. When comparing gene isoforms at the same location but on the different genomes A and B, the term allelic variants will be used. Using this approach, both allelic variants and paralogs of HSP70 cytoplasmic and luminal isoforms were identified in this trail of spots.

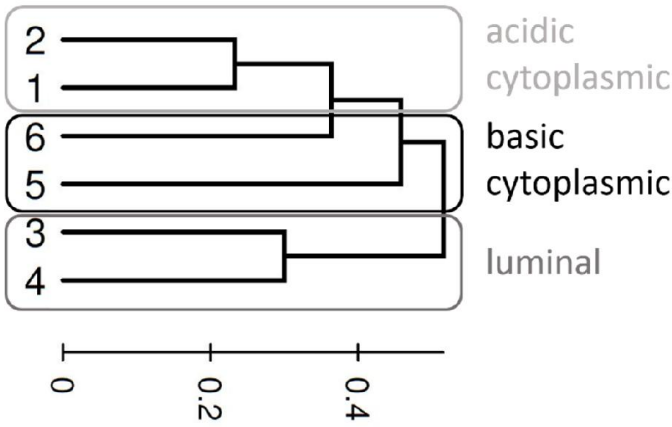

Figure 5.9: Cluster analysis of MALDI-MS spectra of HSP70 spots, represented by their spot number, performed by SPECLUST. Gray boxes indicate the three clusters with the spots which are dominated by the acidic cytoplasmic HSP70 isoforms and the basic cytoplasmic HSP70 isoforms respectively or contain increased amounts of the luminal HSP70 isoforms.

The further characterization of the isoforms has been based on the intensity of a proteotypic peptide to quantify the same isoform over all spots and emPAI scores to quantify the different isoforms within one spot (Table 5.2 and Supplementary file 5.1). The cytoplasmic isoforms are more abundant than the luminal isoforms in all spots as evidenced by higher emPAI scores (Supplementary file 5.1). The isoelectric focusing of one particular isoform is not restricted to one physical location in the gel and each isoform has its highest abundance at a particular isoelectric point (Table 5.2, Figure 5.3 and Supplementary file 5.2). The most abundant spot is spot number 2 with an experimental pI of 5.00 (Figure 5.4). The most abundant paralog in this spot is the chromosome 7 paralog with an emPAI score of 1.94 (7T15160/7_G19958, pI 5.08) (Supplementary file 5.1). Our main spot of interest, spot 5 (pI 5.15), predominantly consists of the chromosome 2 paralog, which has emPAI scores of respectively 0.43 and 0.5 for the A and B allelic variants (pI 5.21) (Supplementary file 5.1). A complete overview of all HSP70 isoforms identified and quantified in each spot is found in Supplementary file 5.1 and Supplementary file 5.2..

Table 5.2: Proteotypic peptides for all cytoplasmic and luminal isoforms with their signature m/z at the MS and MS/MS level.

Paralog [1] (chromosome number)	Allelic variant (sequence id)	Theoretical pI of protein	Peptide sequence of proteotypic peptide	Charge	Theoretical m/z	Y9/Y10/ Y11/Y12	Experimental pI	Spot number	Best ion score
2	A: 2T16250	5.21	MYQGAGGGMGGGMDEDIPSTGGSSGAGPK	3	893.032	1002.486	5.15	5	41.06
	B: 2_G04523	5.21	MYQGAGGGTGGGMDEDIPSTGGSSGAGPK	3	877.703	1002.486	5.15	5	53.70
6	A: 6T34210	5.21	MYQGAGADMAGGMDEDGPTTGGSSAGPK*	3	888.688	959.480	5.21	6	105.68
	B: 6_G18289	5.21							
7	A: 7T15160	5.08	MYQGAGADMAGGMDDDAPPAGGSGAGPK*	3	867.349	895.464	5.00	2	75.51
	B: 7_G19958	5.08							
8	A: 8T20830	5.08	MYQGAGADMGGGMDDDAPSAGASGAGPK	3	864.009	899.459	5.00	2	85.12
	B: 8_G23861	5.05	MYQGAGADMGGEMDDDAPSAGASGAGPK	3	888.016	899.459	4.92	1	65.03
9	A: 9T03960	5.39	MYQGAGADMAGR	3	420.510	n/a	n/a	n/a	n/a
	B: 9_G25413	5.44	MYQGAGADMAGGMDDDVPSAGGSGTGPK	3	883.356	915.454	5.11	4	29.05
10	A: 10T00900	5.06	MYQGAGADMAGEMDDDVPAAGGSGAGPK	3	892.028	869.448	5.06	3	81.77
	B: 10G_28483	5.09	MYQGAGADMAGGMDDDVPSAGGSGAGPK	3	873.353	885.443	5.06	3	79.48
5	A: 5T04580	5.27	DAVVTVPAYFNDSQR*	2	841.413	1097.502	5	2	64.20
	B: 5_G12064	5.42					5.21	6	49.29
3L	A: 3T27250	5.05	SGGAPGGGSDGGDDDDDAHDEL	2	1008.370	n/a	n/a	n/a	n/a
	B: 3_G07942	5.07	SGGAPGGGSDVGDDDDAHDEL	2	971.88	1101.397	5.00	2	59.76
4L	A: 4T11940	5.14	SGGAPGGSDGGDEDDSHDEL	2	937.351	1188.429	5.11	4	43.49
	B: 4_G09464	5.14	SGGSPGGSDGGDEDDSHDEL	2	945.349	1188.429	5.11	4	46.17
9L	A: 9T23710	5.11	SGGAPGGGADGGDDDDSHDEL*	2	950.857	1174.414	5.11	4	49.38
	B: 9_G27669	5.14							

Y9/Y10/Y11/Y12: m/z of the signature y fragment ion conclusively identifying the proteotypic peptide. Best ion score: highest ion score of the proteotypic peptide (Supplementary file 5.1). *Proteotypic peptide at the paralog level. [1]Luminal HSP70s are indicated with an L.

5.3.3 Ubiquitination site prediction and promoter analysis

The abundance of a protein is determined by the rate of protein synthesis and protein breakdown. To get an insight into the different protein isoforms and the differential gene expression and potential protein breakdown, the different protein isoforms were aligned (Supplementary file 5.3), possible ubiquitination sites were predicted and the promoter region of the different genes was analyzed.

Analysis of the predicted ubiquitination pattern by UbiPred and UbPred respectively did not aid in the explanation of differences in protein abundance of different paralogs (Supplementary file 5.4). Moreover, when the two ubiquitination site prediction softwares were compared to each other, it became clear that they return very different results. Only some potential sites were predicted by both softwares which raises the question which software to rely on and prevents meaningful analysis (Figure 5.10).

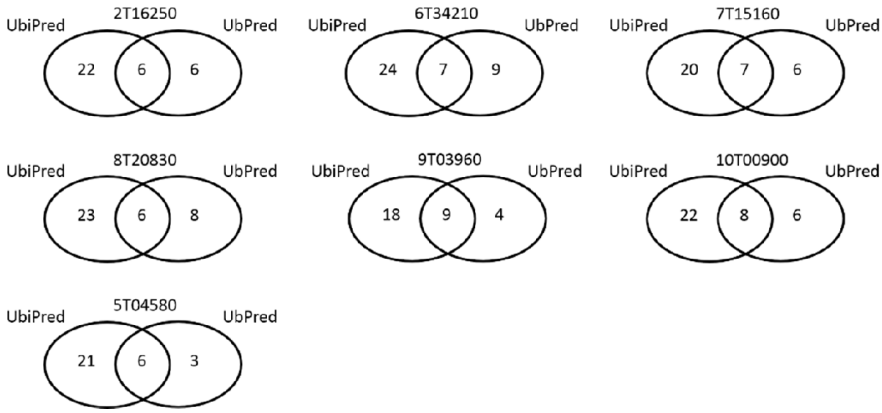

Figure 5.10: Number of potential ubiquitination sites as predicted by UbiPred and UbPred respectively. The intersection contains the number of sites which were predicted by both softwares as potential ubiquitination sites.

Promoter analysis could provide an insight as to why isoforms are expressed with differential abundances and why a specific isoform reacts to certain stressors. We analyzed the promoters of all the identified HSP70s using PLACE and PlantCARE (Higo et al., 1999; Lescot et al., 2002). We focused on the presence of ABA-responsive elements (ABRE), drought responsive elements (DRE) and heat shock elements (HSE) (Figure 5.11). Allelic variants show very similar patterns whereas the different paralogs differ greatly in the number and location of the analyzed elements.

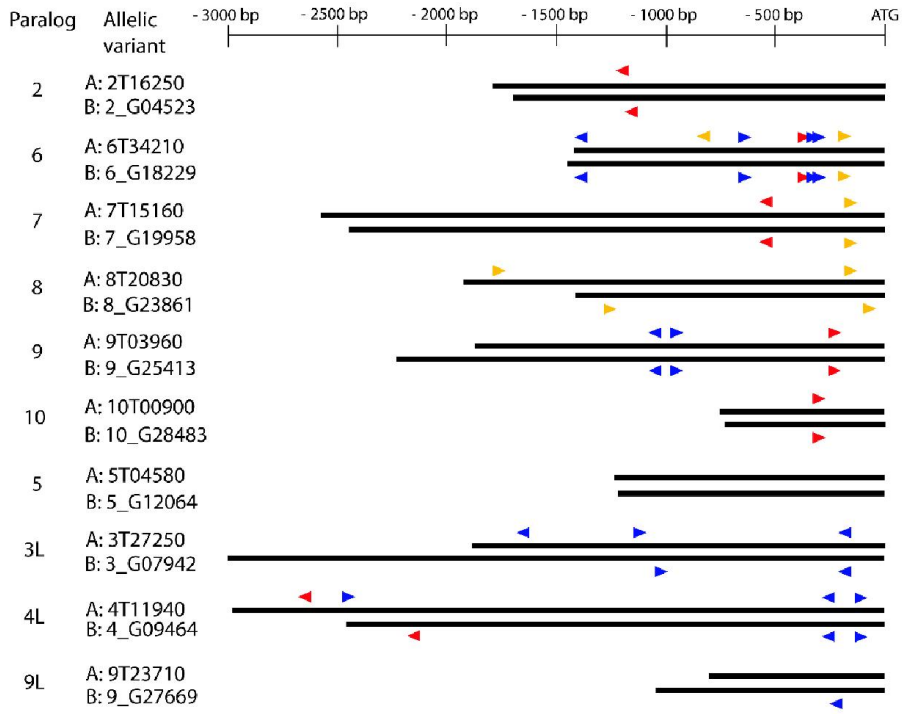

Figure 5.11: Promoter analysis of all identified paralogs and their allelic variants for the cytoplasmic and luminal HSP70. The luminal HSP70 paralogs are indicated with L. Approximate positions and directions of ABRE (ACGTG/TC), DRE (A/CGGAC) and HSE (CNNGAANNTTCNNG) are indicated by red, blue and yellow arrows respectively. The positions and directions are relative to the start codon ATG and up to 3000 bp upstream of this start codon.

5.4 Discussion

5.4.1 Proteomics in a polyploid non-model crop: genomic resources and technical advances

Most identifications of proteins in non-model plants still remain at the level of a gene family when only cross-species identification and or EST databases are available. From our data it is clear that the different paralogs have evolved over time and do not necessarily behave the same when subjected to a stress treatment. Many classical 2DE experiments pick only the differentially expressed spots for identification, ignoring possible isoforms. A functional annotation in a non-model crop is challenging and mostly one is not able to go beyond the gene family annotation (Vincent et al., 2007; Yoshimura et al., 2008; Kjellsen et al., 2010). Even

when a hit from the same organism is found, several spots in a trail can have exactly the same identification and accession number (Taylor et al., 2005). Similarly in *Musa* a trail of HSP70 spots was identified but it was not possible to explore the differences between the spots in the HSP70 trail due to limitations of the genomic/transcriptomic resources at that time (Carpentier et al., 2007; Carpentier et al., 2010). In 2007, only the cross species identification was successful as the available EST database was too limited and did not return any hits. By 2010, a species-specific EST library was produced by our research group but all the HSP70 spots were identified as the same contig. With an A and B genome now available for *Musa*, we have shown that it is possible to identify different protein species at both the paralog and even allelic level. However, the correct structural annotation of the genome is essential in this analysis. We therefore manually curated the necessary HSP70 sequences of the A and B genome. Aside from the genomic resources, the technique for the proteomics analysis is also a deciding factor in the successful identification. A completely gel-free approach was not sufficient to identify which cytoplasmic HSP70 proteins were present in the roots of Cachaco (results not shown). Even with a fast mass spectrometer and long columns (50cm) and gradients, it was impossible to conclusively identify which cytoplasmic HSP70 proteins were present since no tryptic specific peptides were identified. Blind high-throughput gel-free proteomics complicates the study of isoforms of the same gene family. The whole proteome is digested and analyzed in one shot and tryptic specific peptides are likely to go undetected since the common peptides of a gene family will be much more abundant and therefore more likely to be selected for MS/MS analysis. A gel-based approach separates the different HSP70 isoforms based on their pI and mass first before protein identifications and points towards the different abundances of the different isoforms (spots). The standard MALDI-TOF/TOF MS approach still proved insufficient to conclusively identify the isoforms present in each spot as not enough tryptic specific peptides were measured (results not shown). We therefore chose to further separate the digested peptides from the HSP70 spots using liquid chromatography integrating both gel-based and gel-free methods. This further simplifies the mixture and concentrates the peptide that is injected in the mass spectrometer and as a result more peptides were measured. The advantage of an LC-separation is nicely illustrated by the separation of the peptides FSDSSVQSDIK (encoded by gene GSMUA_Achr7T15160) and YSDASVQSDIK (encoded by gene GSMUA_Achr10T00900). These two isoforms of the peptide have the same monoisotopic mass but have different retention times on the RP column because of their different hydrophobicity (approximately 19 and 17 minutes) (Figure 5.12). This would have resulted in a chimeric spectrum using MALDI-TOF/TOF MS but produces separate spectra using LC-MS/MS.

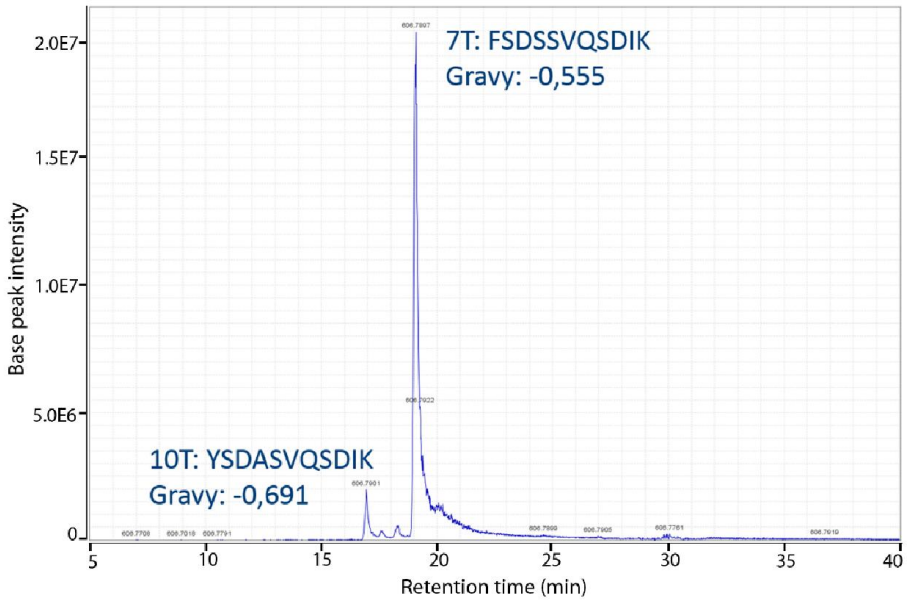

Figure 5.12: Differential retention time illustration of tryptic specific peptides from the HSP70 paralogs of chromosome 10 (YSDASVQSDIK) and of chromosome 7 (FSDSSVQSDIK). Although the peptides share an identical m/z (606.7908), their different amino acid constitution lead to different hydrophobic behavior as evidenced by the different GRAVY scores and therefore different retention times. GRAVY stands for grand average of hydropathicity and peptides with a more negative score are more hydrophilic and are eluted earlier during RP chromatography.

Aside from the added separation and concentration of the peptides provided by the LC, the use of Orbitrap technology significantly adds to the accurate measurement of peptides both at the MS and the MS/MS level (average observed Delta ppm on parent masses was 0.72). Using the combined 2DE LC-MS/MS approach we were able to measure a proteotypic peptide for each paralog and for most of them even at the allelic level (Table 5.2). Moreover, for all isoforms positively identified in at least one spot, we searched the unidentified MS/MS spectra in each spot in which they were not identified. The ion intensity for an MS/MS spectrum was added to the quantification when the peptide fragment mass corresponded to the proteotypic peptide and a specific signature m/z was identified in the MS/MS spectrum. This manual SRM approach shows the perspectives for a gel-free SRM approach to quantify these specific isoforms under different experimental conditions in the future. Gel-based proteomics does remain a time consuming process and a low-throughput technique and both of these pitfalls can be circumvented after initial identification by a targeted gel-free approach.

5.4.2 The *Musa* HSP70

The main objective of this study was to identify the HSP70 protein species that specifically reacted to osmotic stress and to understand what was particular to this protein species. To get an insight into this, the other protein species needed to be characterized as well. The trail of spots contained both cytoplasmic and luminal members of the HSP70 family but the cytoplasmic HSP70s were more abundant.

Structurally HSP70s have two major functional domains, an N-terminal ATPase domain and a peptide binding domain in the C-terminal portion of the protein, connected by a small interdomain hinge. A small C-terminal subdomain is necessary for several co-chaperone interactions. Motifs at the C-terminus in this subdomain can be used to distinguish the subcellular localization of the specific HSP70 protein and are considered a typical plant feature (Sung et al., 2001a). This feature is not completely conserved in all plants, however, as evidenced in rice where certain HSP70 sequences contain these motifs and others do not (Jung et al., 2013).

At the sequence level several HSP70 genes were not correctly structurally annotated on the Banana Genome Hub. Therefore we characterized the whole HSP70 gene family and curated the annotation manually if needed. Out of 20 identified HSP70 gene sequences, 11 sequences belong to our subfamily of interest, the cytoplasmic HSP70s. 14 HSP70 genes were reported in *Arabidopsis thaliana* of which only 5 encode cytoplasmic ones (Lin et al., 2001). Banana has a similar number of HSP70 genes to rice where a total of 24 HSP70 genes were reported of which 11 encode cytoplasmic HSP70s (Sarkar et al., 2013). A typical feature of cytoplasmic isoforms is the EEVD C-terminus. The *Musa* HSP70 family has two members with non-typical termini whereas rice has five so-called non-classical HSP70 genes. Sarkar et al. (2013) suggested that the five non-classical rice genes might be monocot-specific since they were found in rice and sorghum but not in *Arabidopsis*. However our phylogenetic analysis shows that the non-classical *Musa* HSP70s are not similar to the ones in rice (Figure 5.13) therefore the five non-classical rice genes cannot be generalized to be monocot-specific. The cytoplasmic *Musa* HSP70s that do not end on EEVD are 9T03960 and 8T34360 and they do cluster together with rice and *Arabidopsis* HSP70 members and do not form a specific subgroup.

The trail of spots that we identified at the proteome level contains cytoplasmic HSP70s which were also expressed during control conditions. This is not surprising as constitutively expressed HSP70s are known to be key players in protein folding and protein homeostasis (Hartl et al., 2011). But what causes the pronounced increase of spot 5? We had observed this increase in a meristem acclimation study

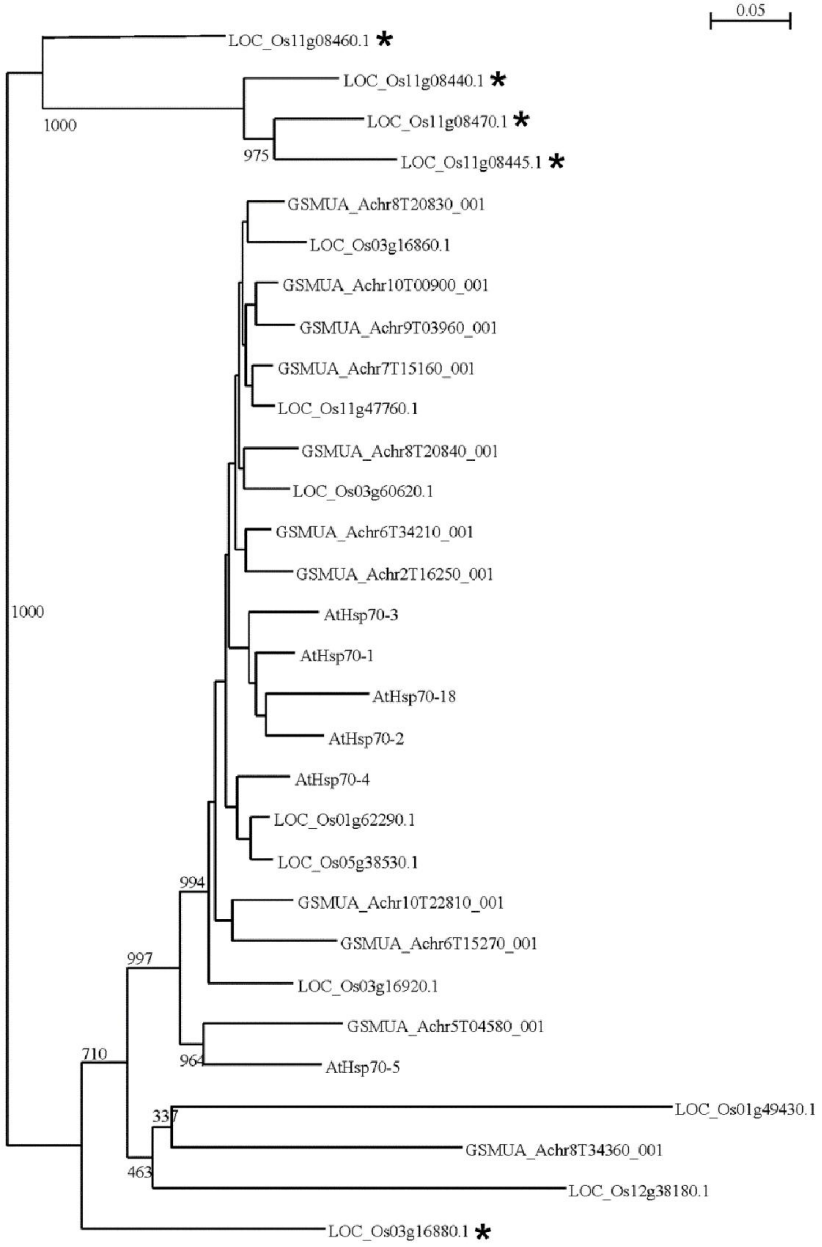

*Figure 5.13: Phylogenetic relationship between all curated cytoplasmic HSP70 protein sequences of the Musa A genome (GSMUA_Achr) and the cytoplasmic HSP70s of rice (LOC_Os) and Arabidopsis (AtHsp). Sequences were aligned using ClustalX and a neighbor-joining tree was constructed with 1000 replicate bootstrap. The bootstrap values between major groups are indicated. The non classical cytoplasmic rice HSP70 accessions have been indicated with *.*

in which this very spot increased the most during sucrose stress (Carpentier et al., 2010) and confirmed this increase in this study (Figure 5.5 and Figure 5.6). To see whether this was a tissue-specific phenomenon and to be able to understand the role HSP70 in roots, we optimized the osmotic concentration to apply moderate stress during which plants suffered a reduction in growth but not a full growth stop. The only HSP70 spot to show an increase in abundance was once again spot 5 (Figure 5.7 and Figure 5.8). The increase was more moderate than during the sucrose stress in meristems but this was not unexpected as the concentration of the sorbitol stress, 0.21M, in the plant experiment was lower than the sucrose concentration during stress in the meristem experiments. In the 2010 study it was also clear that the sucrose stress had a bigger influence on the abundance of this spot than the sorbitol stress. While the whole trail of spots was already known to belong to the HSP70 gene family, we were finally able to identify the isoforms present in the different spots (Figure 5.3).

Since we analyzed the variety Cachaco, a triploid with an unsequenced ABB genome constitution and since it is suggested that the B genome might contribute towards drought tolerance (Simmonds, 1966; Thomas et al., 1998; Robinson and Sauco, 2010), we dug deeper into the expression of the allelic variants from the different genomes A and B. Allelic variants are not necessarily expressed equally as silencing or a biased expression of allelic variants is important in polyploid evolution (Adams et al., 2003; Adams and Wendel, 2005). We have already observed that the proteome phenotype in *Musa* does not always correspond to the expected genome formula due to either gene deletion or gene silencing (De Langhe et al., 2010; Carpentier et al., 2011; Henry et al., 2011). In theory, per locus three different allelic variants could be expressed in Cachaco. Analyzing exactly which allele is expressed and to which amount to which amount it is translated into proteins remains impossible for some loci. The chromosome 5, 6 and 7 allelic variants encoding cytoplasmic HSP70 proteins as well as the chromosome 9 allelic variants encoding a luminal HSP70 protein cannot be discerned from one another (Table 5.2). Although the protein sequences of the allelic variants are not completely identical, the observable tryptic specific peptides are identical. The chromosome 7 genes on the other hand are 100% identical at the protein level (Supplementary file 5.3). No expression was detected of the cytoplasmic chromosome 9 and luminal chromosome 3 allelic variants of the A genome (Supplementary file 5.2). For the chromosome 2, 4, 8 and 10 allelic variants it is clear by analyzing the intensities of the proteotypic peptide that the B allelic variants are more abundantly expressed conform their ABB genome constitution, but the A-specific allelic variant is also expressed.

We determined that the spot with the highest intensity, spot 2, was predominantly made up of the chromosome 7-encoded isoforms. We also identified the chromosome 2-encoded isoforms as the main isoforms in spot 5 which becomes more abundant during osmotic stress. It was already observed before that although members of the HSP70 gene family have a conserved structure and therefore action mechanism, the family is complex and the individual members do play distinct roles (Miernyk, 1997; Sung et al., 2001b; Jung et al., 2013). Our observations corroborate these data. To get an insight into this differential gene expression we first performed a sequence alignment of the main cytoplasmic protein sequences. The alignment showed that only the hinge region and 40 to 50 bp on either side of this region are completely conserved (Supplementary File 5.3). Since several mutations are present in both the N-terminal ATPase domain and the C-terminal peptide binding domain and lid, differences in amino acids are impossible to correlate with differential functions of the isoforms. While a knock-out of the isoforms and subsequent analysis of the phenotypic changes might provide clues as to the functioning of the isoforms, it has been shown in the rice HSP70 family that functional redundancy is present and other HSP70 isoforms take over the role of the knocked-out isoform (Jung et al., 2013).

We therefore analyzed the promoter sequences of all identified HSP70 paralogs and allelic variants for the presence of ABREs, DREs and HSEs (Figure 5.11). It is clear that different paralogs contain different numbers of promoter elements and that these elements are located at different positions in the promoter region. A HSE element however is consistently located within the first 250 bp in the chromosome 6, 7 and 8 paralogs while absent from the other paralogs.

The chromosome 2 allelic variants contained an ABRE element in their promoters at a different location than the other isoforms and at the same time no predicted DRE elements were identified suggesting that the osmotic stress-specific response of spot 5 is most likely ABA dependent (Figure 5.11). Although other promoters also contained ABRE elements, the position of this ABRE element might play a role in the higher abundance of this HSP70 paralog. No HSEs were predicted in the chromosome 2 paralog. HSEs, together with the heat shock transcription factors which bind to them, mediate the heat shock induction. This suggests that the chromosome 2 paralog plays an abscisic acid-mediated stress-specific role in the plant.

The promoter element patterns of the allelic variants, although not identical, revealed similar profiles. This leads us to conclude that in the *Musa* HSP70 family the allelic variants most likely perform similar roles in the plant while the different paralogs contribute to the subfunctionalization.

5.5 Conclusions

In conclusion, this is to our knowledge the first time that a proteomics approach has led to the exploration of protein family at the paralog and allelic variant level in a crop. Moreover we identified a specific osmotic responsive cytoplasmic HSP70 isoform, the HSP70 paralog 2, at the protein level. We have shown that the availability of genomic resources as well as the technique used for proteomics analysis are crucial to go beyond gene family identification. Further research is needed to study the HSP70s during other stresses and to elucidate the roles of the specific isoforms. Additionally, we now have an optimized method which could be applied to other gene families which show interesting dynamics during osmotic or other stresses.

Chapter 6

General conclusions and future perspectives

6.1 General conclusions

This dissertation aimed to further unravel osmotic stress responses in *Musa*, a non-model crop. Our main goals were to validate osmotic candidate stress markers from a meristematic cell model and to investigate mild osmotic stress at the plant level using both heterotrophic and autotrophic models.

Chapter 3 focused on the validation of stress markers identified in earlier acclimation research on the cell model and used a qPCR approach to analyze transcript levels over time. We concluded that the candidate markers PR10, SUMO-conjugating enzyme and ABA-responsive protein react more to the wounding and transfer to new medium than to the applied osmotic stress. One candidate, phosphoglycerate kinase, was confirmed as an osmotic stress marker and can be tested in different tissues in further research. At the time when this research was performed, the *Musa* genome was still unsequenced and readily available sequence data was limited to public and in-house EST databases. Primers were designed based on the EST database, limiting the specificity of the primers. Most of the candidate genes are not single copy genes but belong to a gene family. The phosphoglycerate kinase EST used for primer design for instance has six closely related sequences on the A genome and similarly six allelic variants on the B genome. The primers most likely amplified more than one gene family member. The availability of the A and B genomes however will henceforth make it possible to design gene-specific primers. While it is possible that not all paralogs and/or alleles can be resolved, specific primers will lead to more accurate results.

From the next chapter onward, we switched to the plant level and focused on mild osmotic stress, a stress which causes a growth reduction in the plant but not a growth stop. A long-term set-up using heterotrophic and autotrophic *in vitro* plants with osmotic stress followed by drought experiments on greenhouse plants and a final validation in the field will allow us to identify drought-tolerant varieties.

The heterotrophic *in vitro* growth model allowed us to compare varieties with different genomic constitutions (AAA, AAAh, AAB, AABp and ABB). The ABB Cachaco variety showed the smallest growth reduction during osmotic stress, reconfirming that the *Musa* B genome might be correlated with a higher drought tolerance. A subsequent gel-based proteome analysis of these osmotically stressed and control Cachaco plants revealed that proteins involved in energy metabolism as well as stress and reactive oxygen species metabolism played an important role in reaching a new homeostasis. We used a 2DE approach because protein identification by MS/MS is much simpler in a non-model species due to the reduced protein complexity in a spot. A gel-free approach could be considered from now on for these

differential proteomics studies as the A en B genomes and additional RNA-seq are now available. However 2DE still has a role to play as we showed in the last chapter.

In our gel-based proteomics studies we often encountered the same stress markers. Out of the twenty-three proteins identified in our heterotrophic *in vitro* model, five had been identified previously in proteomics experiments on the meristematic cell model (Table 6.1). Further proteomic research on autotrophic plants again revealed 35 differential proteins of which six had been identified in the previous models (Table 6.1). All these usual suspects are potential stress markers and will be evaluated in the future.

Table 6.1: Usual suspects as identified in the different growth models

Protein name	Meristematic cell model	Heterotrophic *in vitro* plant model	Autotropic plant model
HSP20		x	x
HSP70	x		x
PR10	x	x	
Isoflavone reductase	x	x	
Glutathione-S-transferase	x	x	x
S-adenosyl methionine synthase	x	x	x
Phosphoglucomutase	x	x	
Sucrose synthase	x		x
Phosphoglycerate kinase	x		x

Some of these stress markers, such as HSP70, were however present in more than one spot and not all of these spots show the same response to stress. Proteomics studies on non-model plants can usually not go beyond a gene family identification due to limited sequence resources, but since the A and B genomes of banana became available we successfully went down to the isoform level. However, a manual curation of sequences in both A and B genome was still needed. Only one out of twenty HSP70 sequences present in the A genome was structurally correct. This resulted in an overrepresentation of this one HSP70 isoform in our first analysis. We therefore recommend to check the structural annotation of all genes of a gene family before looking into isoforms. We showed that not only the availability of accurate sequence data but also the right proteomic technology was imperative to reach our goal. 2DE offers the advantage of separating the HSP70, which all have a similar mass of 70kDa, based on their pI. This greatly reduces the complexity in a spot and

although we showed that several isoforms are still present within one spot, we could identify the main isoform based on the ion intensity of the tryptic specific peptides. The 2DE approach was combined with LC-MS/MS. The extra chromatography step further separated peptides after spot digestion and was coupled to a highly accurate mass spectrometer for optimal identification results. The combination of all these techniques allowed us to identify an osmotic responsive HSP70 isoform, the HSP70 paralog 2.

We have now identified and validated nine stress markers. This combination of known stress-correlated genes describes the general osmotic stress response and status of the banana cells. Moreover, from a large gene family (HSP70) we correlated some members to an osmotic stress response. This high level of annotation cannot simply be extrapolated from model organisms. The developed workflow can now be used on other candidates identified in *Musa* and other crops.

6.2 Future perspectives for abiotic stress research

6.2.1 General perspectives

We suggest the following workflow for abiotic stress research in crops based on all the available omics approaches as discussed in chapter 1 and 2 and based on the research performed in this thesis (Figure 6.1).

We propose to start with blind high-throughput analysis using several omics approaches.

(i) Phenomics: We propose to perform phenomics to avoid blind sampling. We suggest to develop methods to select homogenous groups of plants before treatments are applied and/or plants at same stress level rather than at same time point are taken for further analysis. This phenotyping step should pinpoint the moments to be chosen for RNA-seq and proteomic and metabolomic analysis.
(ii) Genomics: The presence of a reference genome is recommended for integration in other omics as a reference genome can be used for anchoring RNA-seq reads and provides a lot of sequence data for protein databases. However the cost of a highly accurate reference genome as well as time needed for assembly and annotation still prevent in practice the sequencing of a reference genome for every crop.
(iii) Transcriptomics: RNA-seq provides vital variety-specific sequence data and expression data. Its integration in protein databases can provide the necessary predicted protein data for unsequenced crops and adds to protein data already

captured in a reference genome as reviewed in chapter 1 and 2. The RNA-seq data can also be used to correct genome annotation as was described in chapter 1.

(iv) Proteomics: We suggest exploratory 2DE for all new crops, varieties and tissues. 2DE is an efficient starting point and allows the easy visualization of isoform trails. Differential proteins and interesting isoform trails remain targets for protein identification. In general, however, we suspect that LC-MS/MS or peptide-based proteomic approaches will become the standard in blind high-throughput differential proteomics. Once again proteomics data can be used to re-annotate certain parts of the genome as we did in chapter 5. Specific workflows will need to be developed as well for low abundant proteins, membrane proteins, isoforms and proteins carrying PTMs as discussed in chapter 1 and 5.

(v) Metabolomics: Like in proteomics, it is still impossible to capture the whole metabolome using one technique. We therefore suggest to also take into account data amassed from proteomics and transcriptomics to analyze specific target metabolites in addition to a wider screen.

In a next phase, all of these omics approaches need to be integrated into one analysis to select targets. As discussed in chapter 1, this integration is still challenging. An integration that enables to represent the data in a comprehensible way will be crucial to identify targets for further analysis.

The last phase includes a targeted analysis to measure the levels of a limited amount of targets in multiple varieties and under several time points. We suggest the use of SRM/MRM at the protein level and qPCR at the transcript level. While requiring some preparation before analysis for each specific target, they can be used on many samples afterwards due to lower cost and time investment. These levels should be correlated to phenotype data as well. To screen multiple varieties under several conditions and/or time points, the use of more automated phenotype screening or phenomics might still be necessary.

While this might be the ideal set-up, it is very clear that this research, even on one crop, cannot be performed by one research group. The cost of such a project as well the expertise required, calls for more cooperation. There is a need for biologists, chemists, (bio-)engineers, bio-informaticians, statisticians, … to all work together on different parts of this puzzle but the main challenge remains to put the pieces together through the exchange and integration of data. While researchers are usually very well equipped to deliver a piece of the puzzle, the integration of all approaches is still rarely seen. Therefore our group has initiated a European network COAST FA1036 (http://www.cost.eu/domains_actions/fa/Actions/FA1306) and has

joined the Flemish initiative for plant phenotyping “Phenovision”. The Hercules project "PHENOVISION" envisages the acquisition and use of an automated phenotyping platform for plants. The project was submitted jointly by UGent, VIB, Hogeschool Gent, KULeuven en UHasselt. The infrastructure is hosted by the VIB in Ghent.

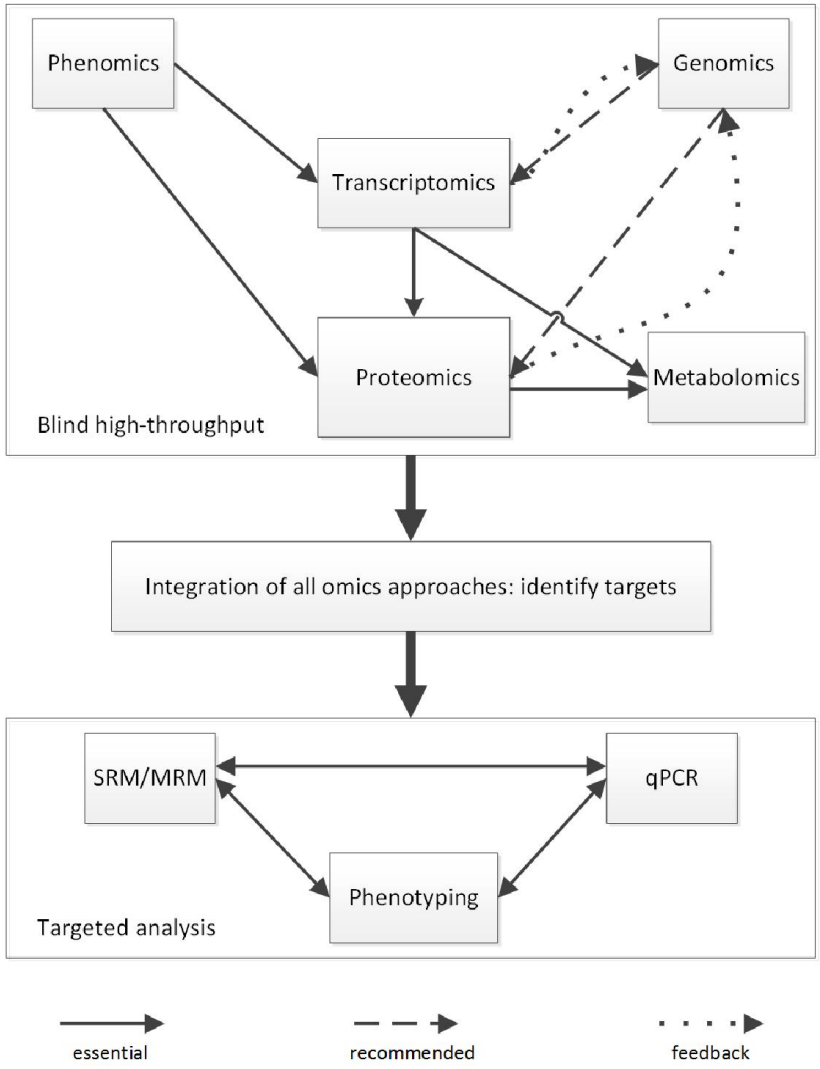

Figure 6.1: Proposed workflow for abiotic stress research in crops. Further clarification can be found in the text.

6.2.2 Perspectives for proteomics research

At the moment, more than 70 crops have been sequenced, but many still remain to be sequenced. On the other hand some crops such as rice and barley and the model plant *Arabidopsis* have already been resequenced. Proteomics is largely based on sequence information for the identification of the proteins. This sequence information is preferably from the same organism but due to the higher conservation of amino acid sequence cross-species identification is possible. However we do believe that RNA-seq is the best first step nowadays in unsequenced organisms. It provides both sequence data as well as quantitative expression data and we therefore recommend to apply it to the studied varieties for both control and at least one stress condition. For sequenced crops for which the genomes of different varieties have not been sequenced yet, it provides variety-specific information which can be integrated in the protein databases as well.

The workflows we suggest depending on sequence status and RNA-seq availability can be found in Figure 6.2. It includes both the set-up of the protein sequence databases as well as the steps to perform with the goal of identifying as many spectra/proteins as possible.

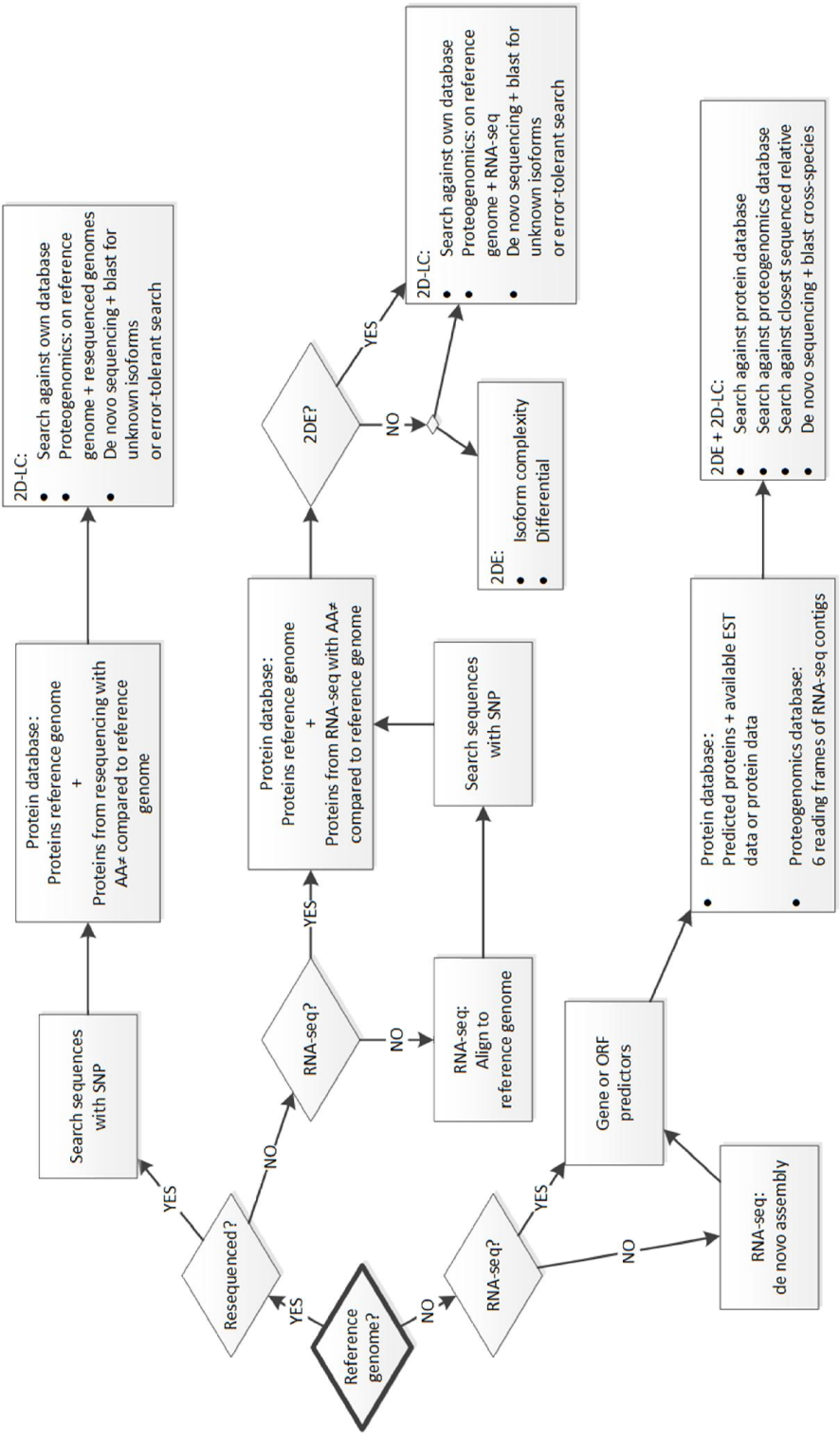

Figure 6.2: Proposed proteomics workflow for crops.

6.3 Future perspectives for osmotic and drought stress research in *Musa*

6.3.1 Future perspectives for proteomics research

Future proteomics research in *Musa* will probably still integrate 2DE, especially for the analysis of certain isoforms. The protein sucrose synthase for instance is present in a trail which contains more than ten spots and many of the sucrose synthase proteins show differential responses to osmotic stress or ABA (Carpentier et al., 2007; Carpentier et al., 2010) (Suzana Garcia, personal communication). Earlier research has already pointed out that isoforms due to PTMs are involved in this trail (unpublished results). However paralogs and allelic variants will most likely also contribute to the trail as eight sucrose synthase sequences have been identified in the A genome. New varieties and tissues will also be analyzed a first time with 2DE as it will allow us to easily spot differences with the varieties and tissues studied before.

For differential proteomics studies, a gel-free approach will be used more as we have already identified the same stress markers several times in our experiments using 2DE. The complementary LC-MS/MS approach then offers a more automated and less time consuming approach while also opening the possibility to analyze proteins not previously covered before. Membrane proteins are also studied using LC-MS/MS as it allows the analysis of the soluble peptides after extraction of the membrane proteins. This approach was already successful in identifying membrane proteins in the unsequenced *Musa* and identification is significantly easier with the availability of the reference genomes (Vertommen et al., 2011a) (Suzana Garcia, personal communication).

6.3.2 Future perspectives for cryopreservation research

We observed that the transcript levels of PR10, SUMO-conjugating enzyme and phosphoglycerate kinase of the osmotically stressed Cachaco meristem cultures had almost returned to their baseline levels after four days. This time point corresponds to the optimal length of acclimation prior to cryopreservation for Cachaco meristems both on sorbitol and sucrose medium. Further research should analyze the transcript levels during sucrose stress, the more suitable sugar for acclimation for cryopreservation, and evaluate whether these transcript levels similarly return to standard levels. Afterwards, other varieties should be investigated to determine whether similar correlations exist between the mRNA levels which return to baseline

readings and the optimal acclimation time point for cryopreservation. We hypothesize that when the stress markers return to the baseline, this indicates full acclimation and that this is the ideal moment to dessicate the cells and freeze them. This research would advance our understanding of cryopreservation survival mechanisms, but has many practical limitations towards application.

6.3.3 Future perspectives for drought stress research

The HSP70 family is being further investigated using qPCR (Tom Scheirs, personal communication) and SRM/MRM. The analysis of the response to several other stresses (e.g. heat and salt) and in different tissues should clarify whether each isoform has its own specific function and whether the isoforms show tissue-specific stress responses.

In a next phase the inclusion of osmotic stress tolerant and sensitive varieties in these HSP70 qPCR and SRM/MRM studies will allow us to screen for differential responses in these varieties and to evaluate the correlation of the HSP70 stress marker to tolerant varieties.

A similar workflow will be used for the other stress markers identified in this research (HSP20, HSP70, PR10, isoflavone reductase, glutathione-S-transferase, S-adenosyl methionine synthase, sucrose synthase, phosphoglucomutase and phosphoglycerate kinase). 2DE will be used to investigate the gene family if multiple protein species have been identified. RNA-seq data will then be used to design paralog- and allele-specific primers while this sequence data can also be used to identify tryptic specific peptides for the SRM/MRM approach.

In a next phase, as proposed in our long-term experimental set-up (Figure 1), the (osmotic) stress markers will be validated as drought stress markers, first in greenhouse plants and later in the field.

Drought stress markers will be evaluated to ascertain their correlation with the severity of drought stress. Once this link has been established, the drought stress marker can be used to precisely monitor stress levels. Combining these drought stress markers with phenotyping approaches could lead to the identification of phenotypes of drought stress which can be evaluated in the field. This concept can then be used to determine when a banana plant has stress and will result in better irrigation management in the field.

The potential of the identified drought stress markers as drought stress tolerance markers will be determined by screening their expression in several drought sensitive and drought tolerant varieties as determined by phenomics approaches.

Once a significant correlation is found, this drought stress tolerance marker can then be used to quickly screen the *Musa* biodiversity present in the *Musa* International Germplasm collection for other tolerant varieties.

A final avenue for future work is the use of this drought tolerance stress markers in classical breeding or genetic engineering. The actual contribution of the marker to drought stress tolerance should be studied using knock-down (RNAi) or overexpression studies. The generation of a more drought stress tolerant plant is unlikely to happen by integrating just one gene, but will most likely involve the simultaneous engineering of multiple genes and/or alleles. In biotic stress, however, monogenic resistance engineering is quite common but therefore not always preferable. To prevent growth reduction during standard conditions, drought-inducible promoters will probably need to be used.

Bibliography

Abreu, I.A., Farinha, A.P., Negrao, S., Goncalves, N., Fonseca, C., Rodrigues, M., Batista, R., Saibo, N.J.M., and Oliveira, M.M. (2013). Coping with abiotic stress: Proteome changes for crop improvement. J Proteomics *93*, 145-168.

Adams, K.L., Cronn, R., Percifield, R., and Wendel, J.F. (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc Natl Acad Sci U S A *100*, 4649-4654.

Adams, K.L., and Wendel, J.F. (2005). Polyploidy and genome evolution in plants. Current Opinion in Plant Biology *8*, 135-141.

Aert, R., Sagi, L., and Volckaert, G. (2004). Gene content and density in banana (Musa Acuminata) as revealed by genomic sequencing of BAC clones. Theor Appl Genet *109*, 129-139.

Agarwal, M., Shrivastava, N., and Padh, H. (2008). Advances in molecular marker techniques and their applications in plant sciences. Plant Cell Reports *27*, 617-631.

Agrawal, G.K., Sarkar, A., Righetti, P.G., Pedreschi, R., Carpentier, S., Wang, T., Barkla, B.J., Kohli, A., Ndimba, B.K., Bykova, N.V.*, et al.* (2013). A decade of plant proteomics and mass spectrometry: Translation of technical advancements to food security and safety issues. Mass Spec Rev *32*, 335-365.

Alban, A., David, S.O., Bjorkesten, L., Andersson, C., Sloge, E., Lewis, S., and Currie, I. (2003). A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. Proteomics *3*, 36-44.

Alm, R., Johansson, P., Hjerno, K., Emanuelsson, C., Ringner, M., and Hakkinen, J. (2006). Detection and identification of protein isoforms using cluster analysis of MALDI-MS mass spectra. J Proteome Res *5*, 785-792.

Anderson, J.V., Li, Q.B., Haskell, D.W., and Guy, C.L. (1994). Structural organization of the spinach endoplasmic reticulum-luminal 70-kilodalton heat-shock cognate gene and expression of 70-kilodalton heat-shock genes during cold acclimation. Plant Physiol *104*, 1359-1370.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis. Nature *408*, 796-815.

Araújo, W.L., Nunes-Nesi, A., and Williams, T.C.R. (2012). Functional genomics tools applied to plant metabolism: a survey on plant respiration, its connections and the annotation of complex gene functions. Frontiers in Plant Science *3*, 210.

Arbona, V., Manzi, M., de Ollas, C., and Gomez-Cadenas, A. (2013). Metabolomics as a tool to investigate abiotic stress tolerance in plants. Int J Mol Sci *14*, 4885-4911.

Babiychuk, E., Kushnir, S., Bellesboix, E., Vanmontagu, M., and Inze, D. (1995). *Arabidopsis thaliana* NADPH oxidoreductase homologs confer tolerance of yeasts towards the thiol-oxidizing drug diamide. Journal of Biological Chemistry *270*, 26224-26231.

Baek, D., Jin, Y., Jeong, J.C., Lee, H.-J., Moon, H., Lee, J., Shin, D., Kang, C.H., Kim, D.H., Nam, J., *et al.* (2008). Suppression of reactive oxygen species by glyceraldehyde-3-phosphate dehydrogenase. Phytochemistry *69*, 333-338.

Baerenfaller, K., Massonnet, C., Walsh, S., Baginsky, S., Buhlmann, P., Hennig, L., Hirsch-Hoffmann, M., Howell, K.A., Kahlau, S., Radziejwoski, A., *et al.* (2012). Systems-based analysis of Arabidopsis leaf growth reveals adaptation to water deficit. Mol Syst Biol *8*, 606.

Bai, T.T., Xie, W.B., Zhou, P.P., Wu, Z.L., Xiao, W.C., Zhou, L., Sun, J., Ruan, X.L., and Li, H.P. (2013). Transcriptome and expression profile analysis of highly resistant and susceptible banana roots challenged with Fusarium oxysporum f. sp. cubense tropical race 4. PloS one *8*, e73945.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007). Quantitative mass spectrometry in proteomics: a critical review. Analytical and Bioanalytical Chemistry *389*, 1017-1031.

Barabaschi, D., Guerra, D., Lacrima, K., Laino, P., Michelotti, V., Urso, S., Vale, G., and Cattivelli, L. (2012). Emerging knowledge from genome sequencing of crop species. Molecular Biotechnology *50*, 250-266.

Barkla, B.J., Vera-Estrella, R., and Pantoja, O. (2013). Progress and challenges for abiotic stress proteomics of crop plants. Proteomics *13*, 1801-1815.

Bartels, D., and Sunkar, R. (2005). Drought and salt tolerance in plants. Critical Reviews in Plant Sciences *24*, 23-58.

Bekele, W.A., Wieckhorst, S., Friedt, W., and Snowdon, R.J. (2013). High-throughput genomics in sorghum: from whole-genome resequencing to a SNP screening array. Plant biotechnology journal *11*, 1112-1125.

Bhardwaj, J., Chauhan, R., Swarnkar, M.K., Chahota, R.K., Singh, A.K., Shankar, R., and Yadav, S.K. (2013). Comprehensive transcriptomic study on horse gram (Macrotyloma uniflorum): De novo assembly, functional characterization and comparative analysis in relation to drought stress. Bmc Genomics *14*, 17.

Bindschedler, L.V., Palmblad, M., and Cramer, R. (2008). Hydroponic isotope labelling of entire plants (HILEP) for quantitative plant proteomics; an oxidative stress case study. Phytochemistry *69*, 1962-1972.

Bindschedler, L.V., and Cramer, R. (2011). Quantitative plant proteomics. Proteomics *11*, 756-775.

Bond, A.E., Row, P.E., and Dudley, E. (2011). Post-translation modification of proteins; methodologies and applications in plant sciences. Phytochemistry *72*, 975-996.

Boschetti, E., Bindschedler, L.V., Tang, C., Fasoli, E., and Righetti, P.G. (2009). Combinatorial peptide ligand libraries and plant proteomics: a winning strategy at a price. Journal of chromatography A *1216*, 1215-1222.

Boston, R.S., Viitanen, P.V., and Vierling, E. (1996). Molecular chaperones and protein folding in plants. Plant Molecular Biology *32*, 191-222.

Braun, R.J., Kinkl, N., Beer, M., and Ueffing, M. (2007). Two-dimensional electrophoresis of membrane proteins. Analytical and Bioanalytical Chemistry *389*, 1033-1045.

Bray, E.A. (1997). Plant responses to water deficit. Trends in Plant Science *2*, 48-54.

Bray, E.A., Bailey-Serres, J., and Weretilnyk, E. (2000). Responses to abiotic stress. In: Biochemistry and molecular biology of plants. Buchanan, B., Gruissem, W., and Jones, R.L., (eds.). American Society of Plant Physiologists, Rockville, Maryland, p. 1158-1203.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S.J., McCurdy, S., Foy, M., Ewan, M.*, et al.* (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nature Biotechnology *18*, 630-634.

Bundock, P.C., Casu, R.E., and Henry, R.J. (2012). Enrichment of genomic DNA for polymorphism detection in a non-model highly polyploid crop plant. Plant biotechnology journal *10*, 657-667.

Buts, K., Michielssens, S., Hertog, M.L., Hayakawa, E., Cordewener, J., America, A.H., Nicolai, B.M., and Carpentier, S.C. (2014). Improving the identification rate of data independent label-free quantitative proteomics experiments on non-model crops: A case study on apple fruit. J Proteomics *105*, 31-45.

Cai, S.G., Wu, D.H., Jabeen, Z., Huang, Y.Q., Huang, Y.C., and Zhang, G.P. (2013). Genome-wide association analysis of aluminum tolerance in cultivated and Tibetan wild barley. PloS one *8*, 11.

Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C.*, et al.* (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nature Genetics *43*, 956-U960.

Carpentier, S., and America, T. (2014). Proteome Analysis of Orphan Plant Species, Fact or Fiction? In: Plant Proteomics. Jorrin-Novo, J.V., Komatsu, S., Weckwerth, W., and Wienkoop, S., (eds.). Humana Press, p. 333-346.

Carpentier, S.C., Witters, E., Laukens, K., Deckers, P., Swennen, R., and Panis, B. (2005). Preparation of protein extracts from recalcitrant plant tissues: An evaluation of different methods for two-dimensional gel electrophoresis analysis. Proteomics *5*, 2497-2507.

Carpentier, S.C., Witters, E., Laukens, K., Van Onckelen, H., Swennen, R., and Panis, B. (2007). Banana (Musa spp.) as a model to study the meristem proteome: Acclimation to osmotic stress. Proteomics *7*, 92-105.

Carpentier, S.C., Coemans, B., Podevin, N., Laukens, K., Witters, E., Matsumura, H., Terauchi, R., Swennen, R., and Panis, B. (2008). Functional genomics in a non-model crop: transcriptomics or proteomics? Physiol Plant *133*, 117-130.

Carpentier, S.C., Swennen, R., and Panis, B. (2009). Plant protein sample preparation for 2DE. In: The protein protocols handbook. Walker, J.M., ed. Humana Press, Totowa, p. 107-117.

Carpentier, S.C., Vertommen, A., Swennen, R., Witters, E., Fortes, C., Souza, M.T., Jr., and Panis, B. (2010). Sugar-mediated acclimation: The importance of sucrose metabolism in meristems. J Proteome Res *9*, 5038-5046.

Carpentier, S.C., Pants, B., Renaut, J., Samyn, B., Vertommen, A., Vanhove, A.-C., Swennen, R., and Sergeant, K. (2011). The use of 2D-electrophoresis and de novo sequencing to characterize inter- and intra-cultivar protein polymorphisms in an allopolyploid crop. Phytochemistry *72*, 1243-1250.

Castellana, N.E., Payne, S.H., Shen, Z.X., Stanke, M., Bafna, V., and Briggs, S.P. (2008). Discovery and revision of Arabidopsis genes by proteogenomics. Proc Natl Acad Sci U S A *105*, 21034-21038.

Castellana, N.E., Shen, Z.X., He, Y.P., Walley, J.W., Cassidy, C.J., Briggs, S.P., and Bafna, V. (2014). An automated proteogenomic method uses mass spectrometry to reveal novel genes in Zea mays. Molecular & Cellular Proteomics *13*, 157-167.

Castro, P.H., Tavares, R.M., Bejarano, E.R., and Azevedo, H. (2012). SUMO, a heavyweight player in plant abiotic stress responses. Cellular and Molecular Life Sciences *69*, 3269-3283.

Champagne, A., and Boutry, M. (2013). Proteomics of nonmodel plant species. Proteomics *13*, 663-673.

Chelius, D., and Bondarenko, P.V. (2002). Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. J Proteome Res *1*, 317-323.

Chene, Y., Rousseau, D., Lucidarme, P., Bertheloot, J., Caffier, V., Morel, P., Belin, E., and Chapeau-Blondeau, F. (2012). On the use of depth camera for 3D phenotyping of entire plants. Computers and Electronics in Agriculture *82*, 122-127.

Coemans, B., Matsumura, H., Terauchi, R., Remy, S., Swennen, R., and Sagi, L. (2005). SuperSAGE combined with PCR walking allows global gene expression profiling of banana (Musa acuminata), a non-model organism. Theor Appl Genet *111*, 1118-1126.

Craig, R., and Beavis, R.C. (2004). TANDEM: matching proteins with tandem mass spectra. Bioinformatics *20*, 1466-1467.

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M.*, et al.* (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature *488*, 213-217.

Damerval, C., Devienne, D., Zivy, M., and Thiellement, H. (1986). Technical improvements in two-dimensional electrophoresis increase the level of genetic-variation detected in wheat-seedling proteins. Electrophoresis *7*, 52-54.

Davey, M., Gudimella, R., Harikrishna, J.A., Sin, L.W., Khalid, N., and Keulemans, J. (2013). A draft Musa balbisiana genome sequence for molecular genetics in polyploid, inter- and intra-specific Musa hybrids. BMC Genomics *14*, 683.

Davey, M.W., Graham, N.S., Vanholme, B., Swennen, R., May, S.T., and Keulemans, J. (2009). Heterologous oligonucleotide microarrays for transcriptomics in a non-model species; a proof-of-concept study of drought stress in Musa. Bmc Genomics *10*, 19.

De Langhe, E., Hribova, E., Carpentier, S., Dolezel, J., and Swennen, R. (2010). Did backcrossing contribute to the origin of hybrid edible bananas? Ann Bot *106*, 849-857.

Dhondt, S., Wuyts, N., and Inze, D. (2013). Cell to whole-plant phenotyping: the best is yet to come. Trends in Plant Science *18*, 433-444.

Downes, B., and Vierstra, R.D. (2005). Post-translational regulation in plants employing a diverse set of polypeptide tags. Biochemical Society Transactions *33*, 393-399.

Droc, G., Larivière, D., Guignon, V., Yahiaoui, N., This, D., Garsmeur, O., Dereeper, A., Hamelin, C., Argout, X., Dufayard, J.-F.*, et al.* (2013). The Banana Genome Hub. Database *2013*.

Dugas, D.V., Monaco, M.K., Olsen, A., Klein, R.R., Kumari, S., Ware, D., and Klein, P.E. (2011). Functional annotation of the transcriptome of Sorghum bicolor in response to osmotic stress and abscisic acid. Bmc Genomics *12*, 21.

Edman, P., and Begg, G. (1967). A protein sequenator. Eur J Biochem *1*, 89-91.

Ellis, R.J. (1979). The most abundant protein in the world. Trends in Biochemical Sciences *4*, 241-244.

Eng, J.K., McCormack, A.L., and Yates, J.R. (1994). An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. Journal of the American Society for Mass Spectrometry *5*, 976-989.

Esteve, C., D'Amato, A., Marina, M.L., Garcia, M.C., and Righetti, P.G. (2013). In-depth proteomic analysis of banana (Musa spp.) fruit with combinatorial peptide ligand libraries. Electrophoresis *34*, 207-214.

Eubel, H., Braun, H.P., and Millar, A.H. (2005). Blue-native PAGE in plants: a tool in analysis of protein-protein interactions. Plant Methods *1:11*.

Fan, J., Wang, H., Feng, D., Liu, B., Liu, H., and Wang, J. (2007). Molecular characterization of plantain class I chitinase gene and its expression in response to infection by Gloeosporium musarum cke and massee and other abiotic stimuli. Journal of Biochemistry *142*, 561-570.

FAO (2012). http://faostat3.fao.org/ (accessed on 2014/04/22)

Fasoli, E., D'Amato, A., Kravchuk, A.V., Boschetti, E., Bachi, A., and Righetti, P.G. (2011). Popeye strikes again: The deep proteome of spinach leaves. J Proteomics *74*, 127-136.

Fernandes, H., Michalska, K., Sikorski, M., and Jaskolski, M. (2013). Structural and functional aspects of PR-10 proteins. Febs Journal *280*, 1169-1199.

Fiehn, O. (2002). Metabolomics--the link between genotypes and phenotypes. Plant Mol Biol *48*, 155-171.

Flexas, J., Galmes, J., Ribas-Carbo, M., and Medrano, H. (2005). The Effects of Water Stress on Plant Respiration. In: Plant Respiration. Lambers, H., and Ribas-Carbo, M., (eds.). Springer Netherlands, p. 85-94.

Ford, K.L., Cassin, A., and Bacic, A. (2011). Quantitative proteomic analysis of wheat cultivars with differing drought stress tolerance. Front Plant Sci *2*, 44.

Frohlich, A., Gaupels, F., Sarioglu, H., Holzmeister, C., Spannagl, M., Durner, J., and Lindermayr, C. (2012). Looking deep inside: detection of low-abundance proteins in leaf extracts of Arabidopsis and phloem exudates of pumpkin. Plant Physiol *159*, 902-914.

Fu, Y., Springer, N.M., Gerhardt, D.J., Ying, K., Yeh, C.T., Wu, W., Swanson-Wagner, R., D'Ascenzo, M., Millard, T., Freeberg, L.*, et al.* (2010). Repeat subtraction-mediated sequence capture from a complex genome. Plant J *62*, 898-909.

Furbank, R.T., and Tester, M. (2011). Phenomics - technologies to relieve the phenotyping bottleneck. Trends in Plant Science *16*, 635-644.

Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., and Bryant, S.H. (2004). Open mass spectrometry search algorithm. J Proteome Res *3*, 958-964.

Ghosh, D., and Xu, J. (2014). Abiotic stress responses in plant roots: a proteomics perspective. Frontiers in Plant Science *5*, 13.

Gilar, M., Olivova, P., Daly, A.E., and Gebler, J.C. (2005). Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. Journal of Separation Science *28*, 1694-1703.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H.*, et al.* (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science *296*, 92-100.

Golldack, D., Li, C., Mohan, H., and Probst, N. (2014). Tolerance to drought and salt stress in plants: unraveling the signaling networks. Frontiers in Plant Science *5*, 10.

Granier, C., Aguirrezabal, L., Chenu, K., Cookson, S.J., Dauzat, M., Hamard, P., Thioux, J.J., Rolland, G., Bouchier-Combaud, S., Lebaudy, A*., et al.* (2006). PHENOPSIS, an automated platform for reproducible phenotyping of plant responses to soil water deficit in Arabidopsis thaliana permitted the identification of an accession with low sensitivity to soil water deficit. New Phytologist *169*, 623-635.

Gross, S.M., Martin, J.A., Simpson, J., Abraham-Juarez, M.J., Wang, Z., and Visel, A. (2013). De novo transcriptome assembly of drought tolerant CAM plants, Agave deserti and Agave tequilana. Bmc Genomics *14*, 14.

Gruhler, A., Schulze, W.X., Matthiesen, R., Mann, M., and Jensen, O.N. (2005). Stable isotope labeling of Arabidopsis thaliana cells and quantitative proteomics by mass spectrometry. Molecular & Cellular Proteomics *4*, 1697-1709.

Guy, C.L., and Li, Q.B. (1998). The organization and evolution of the spinach stress 70 molecular chaperone gene family. Plant Cell *10*, 539-556.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. (1999a). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nature Biotechnology *17*, 994-999.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999b). Correlation between Protein and mRNA Abundance in Yeast. Mol Cell Biol *19*, 1720-1730.

Hartl, F.U., Bracher, A., and Hayer-Hartl, M. (2011). Molecular chaperones in protein folding and proteostasis. Nature *475*, 324-332.

Hashimoto, Y., Strain, B.R., Morimoto, T., and Fukuyama, T. (1984). System identification of plant responses in energy conservative greenhouses. Acta Horticulturae *148*, 287-295.

Hellemans, J., Mortier, G., De Paepe, A., Speleman, F., and Vandesompele, J. (2007). qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. Genome Biology *8*, 14.

Helmy, M., Tomita, M., and Ishihama, Y. (2011). OryzaPG-DB: Rice Proteome Database based on Shotgun Proteogenomics. BMC Plant Biology *11*, 9.

Henry, I.M., Carpentier, S.C., Pampurova, S., Van Hoylandt, A., Panis, B., Swennen, R., and Remy, S. (2011). Structure and regulation of the Asr gene family in banana. Planta *234*, 785-798.

Higashi, Y., and Saito, K. (2013). Network analysis for gene discovery in plant-specialized metabolism. Plant Cell Environ *36*, 1597-1606.

Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Research *27*, 297-300.

Hölscher, D., Dhakshinamoorthy, S., Alexandrov, T., Becker, M., Bretschneider, T., Buerkert, A., Crecelius, A.C., De Waele, D., Elsen, A., Heckel, D.G*., et al.* (2013). Phenalenone-type phytoalexins mediate resistance of banana plants (Musa spp.) to the burrowing nematode Radopholus similis. Proceedings of the National Academy of Sciences, doi: 10.1073/pnas.1314168110.

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics *12*, 14.

Hossain, Z., Khatoon, A., and Komatsu, S. (2013). Soybean Proteomics for Unraveling Abiotic Stress Response Mechanism. J Proteome Res *12*, 4670-4684.

Hoth, S., Morgante, M., Sanchez, J.P., Hanafey, M.K., Tingey, S.V., and Chua, N.H. (2002). Genome-wide gene expression profiling in Arabidopsis thaliana reveals new targets of abscisic acid and largely impaired gene regulation in the abi1-1 mutant. J Cell Sci *115*, 4891-4900.

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z*., et al.* (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet *42*, 961-967.

Ingram, J., and Bartels, D. (1996). The molecular basis of dehydration tolerance in plants. Annual Review of Plant Physiology and Plant Molecular Biology *47*, 377-403.

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. Nature *436*, 793-800.

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics *4*, 1265-1272.

Jackson, S.A., Iwata, A., Lee, S.H., Schmutz, J., and Shoemaker, R. (2011). Sequencing crop genomes: approaches and applications. The New phytologist *191*, 915-925.

Jansen, M., Gilmer, F., Biskup, B., Nagel, K.A., Rascher, U., Fischbach, A., Briem, S., Dreissen, G., Tittmann, S., Braun, S*., et al.* (2009). Simultaneous phenotyping of leaf growth and chlorophyll fluorescence via GROWSCREEN FLUORO allows detection of stress tolerance in Arabidopsis thaliana and other rosette plants. Functional Plant Biology *36*, 902-914.

Jiang, S.-Y., Ramamoorthy, R., and Ramachandran, S. (2008). Comparative transcriptional profiling and evolutionary analysis of the GRAM domain family in eukaryotes. Developmental Biology *314*, 418-432.

Jitsaeng, K., and Schneider, B. (2010). Metabolic profiling of Musa acuminata challenged with Sporobolomyces salmonicolor. Phytochem Lett *3*, 84-87.

Jorrin-Novo, J.V., Maldonado, A.M., Echevarria-Zomeno, S., Valledor, L., Castillejo, M.A., Curto, M., Valero, J., Sghaier, B., Donoso, G., and Redondo, I. (2009). Plant proteomics update (2007-2008): Second-generation proteomic techniques, an appropriate experimental design, and data analysis to fulfill MIAPE standards, increase plant proteome coverage and expand biological knowledge. J Proteomics *72*, 285-314.

Jorrín, J.V., Maldonado, A.M., and Castillejo, M.A. (2007). Plant proteome analysis: a 2006 update. proteomics *7*, 2947-2962.

Jung, K.-H., Gho, H.-J., Nguyen, M., Kim, S.-R., and An, G. (2013). Genome-wide expression analysis of HSP70 family genes in rice and identification of a cytosolic HSP70 gene highly induced under heat stress. Funct Integr Genomics *13*, 391-402.

Kakumanu, A., Ambavaram, M.M.R., Klumas, C., Krishnan, A., Batlang, U., Myers, E., Grene, R., and Pereira, A. (2012). Effects of drought on gene expression in maize reproductive and leaf meristem tissue revealed by RNA-seq. Plant Physiol *160*, 846-867.

Kjellsen, T.D., Shiryaeva, L., Schroder, W.P., and Strimbeck, G.R. (2010). Proteomics of extreme freezing tolerance in Siberian spruce (Picea obovata). J Proteomics *73*, 965-975.

Kosova, K., Vitamvas, P., Prasil, I.T., and Renaut, J. (2011). Plant proteome changes under abiotic stress - Contribution of proteomics studies to understanding plant stress response. J Proteomics *74*, 1301-1322.

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z., Song, W., Ying, K., Zhang, M*., et al.* (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet *42*, 1027-1030.

Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.-L., Li, M.-W., He, W., Qin, N., Wang, B., *et al.* (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet *42*, 1053-1059.

Lange, V., Picotti, P., Domon, B., and Aebersold, R. (2008). Selected reaction monitoring for quantitative proteomics: a tutorial. Mol Syst Biol *4*, 222.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., *et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

Laugesen, S., Bak-Jensen, K.S., Hagglund, P., Henriksen, A., Finnie, C., Svensson, B., and Roepstorff, P. (2007). Barley peroxidase isozymes - Expression and post-translational modification in mature seeds as identified by two-dimensional gel electrophoresis and mass spectrometry. International Journal of Mass Spectrometry *268*, 244-253.

Lehmann, U., Wienkoop, S., Tschoep, H., and Weckwerth, W. (2008). If the antibody fails – a mass Western approach. The Plant Journal *55*, 1039-1046.

Leister, D., Varotto, C., Pesaresi, P., Niwergall, A., and Salamini, F. (1999). Large-scale evaluation of plant growth in Arabidopsis thaliana by non-invasive image analysis. Plant Physiol Biochem *37*, 671-678.

Lescot, M., Dehais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., Rouze, P., and Rombauts, S. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Research *30*, 325-327.

Li, C.Q., Shao, J.F., Wang, Y.J., Li, W.B., Guo, D.J., Yan, B., Xia, Y.J., and Peng, M. (2013). Analysis of banana transcriptome and global gene expression profiles in banana roots in response to infection by race 1 and tropical race 4 of Fusarium oxysporum f. sp cubense. Bmc Genomics *14*, 16.

Li, C.Y., Deng, G.M., Yang, J., Viljoen, A., Jin, Y., Kuang, R.B., Zuo, C.W., Lv, Z.C., Yang, Q.S., Sheng, O., *et al.* (2012). Transcriptome profiling of resistant and susceptible Cavendish banana roots following inoculation with Fusarium oxysporum f. sp. cubense tropical race 4. BMC Genomics *13*, 374.

Li, W., Jaroszewski, L., and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics *17*, 282-283.

Lin, B.L., Wang, J.S., Liu, H.C., Chen, R.W., Meyer, Y., Barakat, A., and Delseny, M. (2001). Genomic analysis of the Hsp70 superfamily in Arabidopsis thaliana. Cell Stress & Chaperones *6*, 201-208.

Liu, H.-Y., Dai, J.-R., Feng, D.-R., Liu, B., Wang, H.-B., and Wang, J.-F. (2010). Characterization of a Novel Plantain Asr Gene, MpAsr, that is Regulated in Response to Infection of Fusarium oxysporum f. sp cubense and Abiotic Stresses. Journal of Integrative Plant Biology *52*, 315-323.

Liu, J.-J., and Ekramoddoullah, A.K.M. (2006). The family 10 of plant pathogenesis-related proteins: Their structure, regulation, and function in response to biotic and abiotic stresses. Physiological and Molecular Plant Pathology *68*, 3-13.

Manza, L.L., Stamer, S.L., Ham, A.J.L., Codreanu, S.G., and Liebler, D.C. (2005). Sample preparation and digestion for proteomic analyses using spin filters. Proteomics *5*, 1742-1745.

Marin, D.H., Romero, R.A., Guzman, M., and Sutton, T.B. (2003). Black sigatoka: An increasing threat to banana cultivation. Plant Disease *87*, 208-222.

Matsumura, H., Reich, S., Ito, A., Saitoh, H., Kamoun, S., Winter, P., Kahl, G., Reuter, M., K, D.H., and Terauchi, R. (2003). Gene expression analysis of plant host-pathogen interactions by SuperSAGE. Proc Natl Acad Sci U S A *100*, 15718-15723.

Mechin, V., Damerval, C., and Zivy, M. (2007). Total protein extraction with TCA-acetone. In: Methods Mol Biol. Thiellement, H., M., Z., C., D., and V., M., (eds.). p. 1-8.

Michael, T.P., and Jackson, S. (2013). The first 50 plant genomes. Plant Gen *6*, 10.3835/plantgenome2013.3803.0001in.

Michalski, A., Cox, J., and Mann, M. (2011). More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. J Proteome Res *10*, 1785-1793.

Miernyk, J.A. (1997). The 70 kDa stress-related proteins as molecular chaperones. Trends in Plant Science *2*, 180-187.

Miller, G.A.D., Suzuki, N., Ciftci-Yilmaz, S., and Mittler, R.O.N. (2010). Reactive oxygen species homeostasis and signalling during drought and salinity stresses. Plant, Cell & Environment *33*, 453-467.

Mirzaei, M., Pascovici, D., Atwell, B.J., and Haynes, P.A. (2012a). Differential regulation of aquaporins, small GTPases and V-ATPases proteins in rice leaves subjected to drought stress and recovery. Proteomics *12*, 864-877.

Mirzaei, M., Soltani, N., Sarhadi, E., Pascovici, D., Keighley, T., Salekdeh, G.H., Haynes, P.A., and Atwell, B.J. (2012b). Shotgun proteomic analysis of long-distance drought signaling in rice roots. J Proteome Res *11*, 348-358.

Mitchell-Olds, T. (2010). Complex-trait analysis in plants. Genome Biol *11*, 113.

Mitra, J. (2001). Genetics and genetic improvement of drought resistance in crop plants. Curr Sci *80*, 758-763.

Mittler, R. (2002). Oxidative stress, antioxidants and stress tolerance. Trends in Plant Science *7*, 405-410.

Mittler, R., and Shulaev, V. (2013). Functional genomics, challenges and perspectives for the future. Physiol Plant *148*, 317-321.

Miyakawa, T., Fujita, Y., Yamaguchi-Shinozaki, K., and Tanokura, M. (2013). Structure and function of abscisic acid receptors. Trends in Plant Science *18*, 259-266.

Mock, H.P., Heller, W., Molina, A., Neubohn, B., Sandermann, H., Jr., and Grimm, B. (1999). Expression of uroporphyrinogen decarboxylase or coproporphyrinogen oxidase antisense RNA in tobacco induces pathogen defense responses conferring increased resistance to tobacco mosaic virus. J Biol Chem *274*, 4231-4238.

Moller, I.M. (2001). Plant mitochondria and oxidative stress: Electron transport, NADPH turnover, and metabolism of reactive oxygen species. Annual Review of Plant Physiology and Plant Molecular Biology *52*, 561-591.

Morey, M., Fernandez-Marmiesse, A., Castineiras, D., Fraga, J.M., Couce, M.L., and Cocho, J.A. (2013). A glimpse into past, present, and future DNA sequencing. Mol Genet Metab *110*, 3-24.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods *5*, 621-628.

Nagel, K.A., Kastenholz, B., Jahnke, S., Van Dusschoten, D., Aach, T., Muehlich, M., Truhn, D., Scharr, H., Terjung, S., Walter, A.*, et al.* (2009). Temperature responses of roots: impact on growth, root system architecture and implications for phenotyping. Functional Plant Biology *36*, 947-959.

Nagel, K.A., Putz, A., Gilmer, F., Heinz, K., Fischbach, A., Pfeifer, J., Faget, M., Blossfeld, S., Ernst, M., Dimaki, C*., et al.* (2012). GROWSCREEN-Rhizo is a novel phenotyping robot enabling simultaneous measurements of root and shoot growth for plants grown in soil-filled rhizotrons. Functional Plant Biology *39*, 891-904.

Neilson, K.A., Mariani, M., and Haynes, P.A. (2011). Quantitative proteomic analysis of cold-responsive proteins in rice. Proteomics *11*, 1696-1706.

Neuhoff, V., Arold, N., Taube, D., and Ehrhardt, W. (1988). Improved staining of proteins in polyacrylamide gels including isoelectric-focusing gels with clear background at nanogram sensitivity using coomassie brilliant blue G-250 and R-250. Electrophoresis *9*, 255-262.

Neves, L.G., Davis, J.M., Barbazuk, W.B., and Kirst, M. (2013). Whole-exome targeted sequencing of the uncharacterized pine genome. Plant J *75*, 146-156.

Ngara, R., and Ndimba, B.K. (2014). Understanding the complex nature of salinity and drought-stress response in cereals using proteomics technologies. Proteomics *14*, 611-621.

Novatchkova, M., Tomanov, K., Hofmann, K., Stuible, H.P., and Bachmair, A. (2012). Update on sumoylation: defining core components of the plant SUMO conjugation system by phylogenetic comparison. New Phytologist *195*, 23-31.

O'Farrell, P.H. (1975). High resolution two-dimensional electrophoresis of proteins. J Biol Chem *250*, 4007-4021.

Obata, T., and Fernie, A.R. (2012). The use of metabolomics to dissect plant responses to abiotic stresses. Cellular and Molecular Life Sciences *69*, 3225-3243.

Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, Silac, as a simple and accurate approach to expression proteomics. Molecular & Cellular Proteomics *1*, 376-386.

Oono, Y., Kawahara, Y., Yazawa, T., Kanamori, H., Kuramata, M., Yamagata, H., Hosokawa, S., Minami, H., Ishikawa, S., Wu, J.Z*., et al.* (2013a). Diversity in the complexity of phosphate starvation transcriptomes among rice cultivars based on RNA-Seq profiles. Plant Molecular Biology *83*, 523-537.

Oono, Y., Kobayashi, F., Kawahara, Y., Yazawa, T., Handa, H., Itoh, T., and Matsumoto, T. (2013b). Characterisation of the wheat (triticum aestivum L.) transcriptome by de novo assembly for the discovery of phosphate starvation-responsive genes: gene expression in Pi-stressed wheat. Bmc Genomics *14*, 14.

Otálvaro, F., Nanclares, J., Vásquez, L.E., Quiñones, W., Echeverri, F., Arango, R., and Schneider, B. (2007). Phenalenone-Type Compounds from Musa acuminata var. "Yangambi km 5" (AAA) and Their Activity against Mycosphaerella fijiensis. Journal of Natural Products *70*, 887-890.

Panis, B., Strosse, H., Van Den Hende, S., and Swennen, R. (2002). Sucrose preculture to simplify cryopreservation of banana meristem cultures. Cryoletters *23*, 375-384.

Pardo, J.M. (2010). Biotechnology of water and salinity stress tolerance. Curr Opin Biotechnol *21*, 185-196.

Pariset, L., Chillemi, G., Bongiorni, S., Romano Spica, V., and Valentini, A. (2009). Microarrays and high-throughput transcriptomic analysis in species with incomplete availability of genomic sequences. New biotechnology *25*, 272-279.

Passioura, J. (2007). The drought environment: physical, biological and agricultural perspectives. J Exp Bot *58*, 113-117.

Pellicer, J., Fay, M.F., and Leitch, I.J. (2010). The largest eukaryotic genome of them all? Bot J Linn Soc *164*, 10-15.

Perkins, D.N., Pappin, D.J.C., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis *20*, 3551-3567.

Perrière, G., and Gouy, M. (1996). WWW-query: An on-line retrieval system for biological sequence banks. Biochimie *78*, 364-369.

Plumb, R.S., Johnson, K.A., Rainville, P., Smith, B.W., Wilson, I.D., Castro-Perez, J.M., and Nicholson, J.K. (2006). UPLC/MSE; a new approach for generating molecular fragment information for biomarker structure elucidation. Rapid Communications in Mass Spectrometry *20*, 1989-1994.

Poczai, P., Varga, I., Laos, M., Cseh, A., Bell, N., Valkonen, J., and Hyvonen, J. (2013). Advances in plant gene-targeted and functional markers: a review. Plant Methods *9*, 6.

Podevin, N., Krauss, A., Henry, I., Swennen, R., and Remy, S. (2012). Selection and validation of reference genes for quantitative RT-PCR expression studies of the non-model crop Musa. Mol Breeding *30*, 1237-1252.

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., Sorrells, M*., et al.* (2012a). Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. Plant Gen *5*, 103-113.

Poland, J.A., Brown, P.J., Sorrells, M.E., and Jannink, J.L. (2012b). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PloS one *7*, e32253.

Postnikova, O.A., Shao, J., and Nemchinov, L.G. (2013). Analysis of the alfalfa root transcriptome in response to salinity stress. Plant and Cell Physiology *54*, 1041-1055.

Rabilloud, T., and Chevallet, M. (2000). Solubilization of proteins in two-dimensional electrophoresis. In: Proteome Research: Two-Dimensional Gel Electrophoresis and Identification Methods. Rabilloud, T., ed. Springer Heidelberg, p. 9-29.

Radivojac, P., Vacic, V., Haynes, C., Cocklin, R.R., Mohan, A., Heyen, J.W., Goebl, M.G., and Iakoucheva, L.M. (2010). Identification, analysis, and prediction of protein ubiquitination sites. Proteins-Structure Function and Bioinformatics *78*, 365-380.

Ragupathy, R., You, F.M., and Cloutier, S. (2013). Arguments for standardizing transposable element annotation in plant genomes. Trends in Plant Science *18*, 367-376.

Rampitsch, C., and Bykova, N.V. (2012). The beginnings of crop phosphoproteomics: exploring early warning systems of stress. Frontiers in Plant Science *3*, 144.

Rappsilber, J., Ryder, U., Lamond, A.I., and Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. Genome Res *12*, 1231-1245.

Ravi, I., Uma, S., Vaganan, M.M., and Mustaffa, M.M. (2013). Phenotyping bananas for drought resistance. Frontiers in Physiology *4*, 9.

Reddy, A.R., Chaitanya, K.V., and Vivekanandan, M. (2004). Drought-induced responses of photosynthesis and antioxidant metabolism in higher plants. Journal of Plant Physiology *161*, 1189-1202.

Reese, M.G., and Guigo, R. (2006). EGASP: Introduction. Genome Biology *7*, 3.

Reinartz, J., Bruyns, E., Lin, J.-Z., Burcham, T., Brenner, S., Bowen, B., Kramer, M., and Woychik, R. (2002). Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. Briefings in Functional Genomics & Proteomics *1*, 95-104.

Remmerie, N., De Vijlder, T., Laukens, K., Dang, T.H., Lemiere, F., Mertens, I., Valkenborg, D., Blust, R., and Witters, E. (2011). Next generation functional proteomics in non-model plants: A survey on techniques and applications for the analysis of protein complexes and post-translational modifications. Phytochemistry *72*, 1192-1218.

Renuse, S., Chaerkady, R., and Pandey, A. (2011). Proteogenomics. Proteomics *11*, 620-630.

Righetti, P.G., Boschetti, E., Lomas, L., and Citterio, A. (2006). Protein equalizer (TM) technology: the quest for a "democratic proteome". Proteomics *6*, 3980-3992.

Robinson, J.C., and Sauco, V.G. (2010). Bananas and Plantains, 2nd Edition (Wallingford OX10 8DE, Oxon, UK: CABI Publishing-CAB International).

Rogowska-Wrzesinska, A., Le Bihan, M.-C., Thaysen-Andersen, M., and Roepstorff, P. (2013). 2D gels still have a niche in proteomics. J Proteomics *88*, 4-13.

Rolland, F., Moore, B., and Sheen, J. (2002). Sugar sensing and signaling in plants. Plant Cell *14*, S185-205.

Rouard, M., Guignon, V., Aluome, C., Laporte, M.-A., Droc, G., Walde, C., Zmasek, C.M., Perin, C., and Conte, M.G. (2011). GreenPhylDB v2.0: comparative and functional genomics in plants. Nucleic Acids Research *39*, D1095-D1102.

Roux-Dalvai, F., Gonzalez de Peredo, A., Simo, C., Guerrier, L., Bouyssie, D., Zanella, A., Citterio, A., Burlet-Schiltz, O., Boschetti, E., Righetti, P.G*., et al.* (2008). Extensive analysis of the cytoplasmic proteome of human erythrocytes using the peptide ligand library technology and advanced mass spectrometry. Mol Cell Proteomics *7*, 2254-2269.

Roux, N., Baurens, F.-C., Dolezel, J., Hribova, E., Heslop-Harrison, P., Town, C., Sasaki, T., Matsumoto, T., Aert, R., Remy, S*., et al.* (2008). Genomics of banana and plantain (Musa spp.), major staple crops in the tropics, Vol 1 (Springer, 233 Spring Street, New York, Ny 10013, United States).

Roy, A., Rushton, P.J., and Rohila, J.S. (2011). The potential of proteomics technologies for crop improvement under drought conditions. Critical Reviews in Plant Sciences *30*, 471-490.

Roychoudhury, A., Paul, S., and Basu, S. (2013). Cross-talk between abscisic acid-dependent and abscisic acid-independent pathways during abiotic stress. Plant Cell Reports *32*, 985-1006.

Rukundo, P. (2009). Evaluation of the water use efficiency of different *Musa* varieties: Development of a sorbitol induced osmotic stress in vitro model (Master thesis).

Rukundo, P., Carpentier, S.C., and Swennen, R. (2012). Development of in vitro technique to screen for drought tolerant banana varieties by sorbitol induced osmotic stress. African Journal of Plant Science *6*, 416-425.

Salekdeh, G.H., and Komatsu, S. (2007). Crop proteomics: Aim at sustainable agriculture of tomorrow. Proteomics *7*, 2976-2996.

Saravanan, R.S., and Rose, J.K.C. (2004). A critical evaluation of sample extraction techniques for enhanced proteomic analysis of recalcitrant plant tissues. Proteomics *4*, 2522-2532.

Sardans, J., Peñuelas, J., and Rivas-Ubach, A. (2011). Ecological metabolomics: overview of current developments and future challenges. Chemoecology *21*, 191-225.

Sarkar, N.K., Kundnani, P., and Grover, A. (2013). Functional analysis of Hsp70 superfamily proteins of rice (Oryza sativa). Cell Stress & Chaperones *18*, 427-437.

Schaff, J.E., Mbeunkui, F., Blackburn, K., Bird, D.M., and Goshe, M.B. (2008). SILIP: a novel stable isotope labeling method for in planta quantitative proteomic analysis. The Plant Journal *56*, 840-854.

Schatz, M.C., Witkowski, J., and McCombie, W.R. (2012). Current challenges in de novo plant genome sequencing and assembly. Genome Biology *13*, 7.

Schmidt, A., Kellermann, J., and Lottspeich, F. (2005). A novel strategy for quantitative proteornics using isotope-coded protein labels. Proteomics *5*, 4-15.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J.*, et al.* (2010). Genome sequence of the palaeopolyploid soybean. Nature *463*, 178-183.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F.S., Pasternak, S., Liang, C.Z., Zhang, J.W., Fulton, L., Graves, T.A.*, et al.* (2009). The B73 Maize genome: complexity, diversity, and dynamics. Science *326*, 1112-1115.

Schuster, A.M., and Davies, E. (1983). Ribonucleic-acid and protein-metabolism in Pea epicotyls .1. The aging process. Plant Physiol *73*, 809-816.

Serraj, R., and Sinclair, T.R. (2002). Osmolyte accumulation: can it really help increase crop yield under drought conditions? Plant Cell and Environment *25*, 333-341.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research *13*, 2498-2504.

Shekhawat, U.K., Ganapathi, T.R., and Srinivas, L. (2011a). Cloning and characterization of a novel stress-responsive WRKY transcription factor gene (MusaWRKY71) from Musa spp. cv. Karibale Monthan (ABB group) using transformed banana cells. Mol Biol Rep *38*, 4023-4035.

Shekhawat, U.K., Srinivas, L., and Ganapathi, T.R. (2011b). MusaDHN-1, a novel multiple stress-inducible SK(3)-type dehydrin gene, contributes affirmatively to drought- and salt-stress tolerance in banana. Planta *234*, 915-932.

Shevchenko, A., Tomas, H., Havlis, J., Olsen, J.V., and Mann, M. (2006). In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nature Protocols *1*, 2856-2860.

Shinozaki, K., and Yamaguchi-Shinozaki, K. (2000). Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. Current Opinion in Plant Biology *3*, 217-223.

Shliaha, P.V., Bond, N.J., Gatto, L., and Lilley, K.S. (2013). Effects of traveling wave ion mobility separation on data independent acquisition in proteomics studies. J Proteome Res *12*, 2323-2339.

Sihlbom, C., Kanmert, I., Bahr, H., and Davidsson, P. (2008). Evaluation of the combination of bead technology with SELDI-TOF-MS and 2-D DIGE for detection of plasma proteins. J Proteome Res *7*, 4191-4198.

Simmonds, N.W. (1966). Bananas (London: Longmans).

Skirycz, A., and Inzé, D. (2010). More from less: plant growth under limited water. Current Opinion in Biotechnology *21*, 197-203.

Skirycz, A., Vandenbroucke, K., Clauw, P., Maleux, K., De Meyer, B., Dhondt, S., Pucci, A., Gonzalez, N., Hoeberichts, F., Tognetti, V.B.*, et al.* (2011). Survival and growth of Arabidopsis plants given limited water are not equal. Nature Biotechnology *29*, 212-214.

Sobhanian, H., Aghaei, K., and Komatsu, S. (2011). Changes in the plant proteome resulting from salt stress: Toward the creation of salt-tolerant crops? J Proteomics *74*, 1323-1337.

Soltis, D.E., Soltis, P.S., and Tate, J.A. (2004). Advances in the study of polyploidy since Plant speciation. New Phytologist *161*, 173-191.

Strosse, H., Schoofs, H., Panis, B., Andre, E., Reyniers, K., and Swennen, R. (2006). Development of embryogenic cell suspensions from shoot meristematic tissue in bananas and plantains (Musa spp.). Plant Sci *170*, 104-112.

Stuckens, J., Dzikiti, S., Verstraeten, W.W., Verreynne, S., Swennen, R., and Coppin, P. (2011). Physiological interpretation of a hyperspectral time series in a citrus orchard. Agric For Meteorol *151*, 1002-1015.

Sulpice, R., Trenkamp, S., Steinfath, M., Usadel, B., Gibon, Y., Witucka-Wall, H., Pyl, E.-T., Tschoep, H., Steinhauser, M.C., Guenther, M.*, et al.* (2010). Network Analysis of Enzyme Activities and Metabolite Levels and Their Relationship to Biomass in a Large Panel of Arabidopsis Accessions. The Plant Cell Online *22*, 2872-2893.

Sung, D.Y., Kaplan, F., and Guy, C.L. (2001a). Plant Hsp70 molecular chaperones: Protein structure, gene family, expression and function. Physiol Plant *113*, 443-451.

Sung, D.Y., Vierling, E., and Guy, C.L. (2001b). Comprehensive expression profile analysis of the Arabidopsis Hsp70 gene family. Plant Physiol *126*, 789-800.

Swindell, W.R., Huebner, M., and Weber, A.P. (2007). Transcriptional profiling of Arabidopsis heat shock proteins and transcription factors reveals extensive overlap between heat and non-heat stress response pathways. BMC Genomics *8*.

Taylor, N.L., Heazlewood, J.L., Day, D.A., and Millar, A.H. (2005). Differential impact of environmental stresses on the pea mitochondrial proteome. Molecular & Cellular Proteomics *4*, 1122-1133.

Thomas, D.S., Turner, D., and Eamus, D. (1998). Independent effects of the environment on the leaf gas exchange of three banana (Musa spp.) cultivars of different genomic constitution. Scientia Horticulturae *75*, 41-57.

Thulasiraman, V., Lin, S., Gheorghiu, L., Lathrop, J., Lomas, L., Hammond, D., and Boschetti, E. (2005). Reduction of the concentration difference of proteins in biological liquids using a library of combinatorial ligands. Electrophoresis *26*, 3561-3571.

Tung, C.-W., and Ho, S.-Y. (2008). Computational identification of ubiquitylation sites from protein sequences. BMC Bioinformatics *9*.

Valdés, A., Ibanez, C., Simo, C., and Garcia-Canas, V. (2013). Recent transcriptomics advances and emerging applications in food science. Trac-Trends in Analytical Chemistry *52*, 142-154.

Valentine, S.J., Kulchania, M., Barnes, C.A.S., and Clemmer, D.E. (2001). Multidimensional separations of complex peptide mixtures: a combined high-performance liquid chromatography/ion mobility/time-of-flight mass spectrometry approach. International Journal of Mass Spectrometry *212*, 97-109.

van Asten, P.J.A., Fermont, A.M., and Taulya, G. (2011). Drought is a major yield loss factor for rainfed East African highland banana. Agric Water Manage *98*, 541-552.

van der Heijden, G., Song, Y., Horgan, G., Polder, G., Dieleman, A., Bink, M., Palloix, A., van Eeuwijk, F., and Glasbey, C. (2012). SPICY: towards automated phenotyping of large pepper plants in the greenhouse. Functional Plant Biology *39*, 870-877.

van Loon, L.C., and Van Strien, E.A. (1999). The families of pathogenesis-related proteins, their activities, and comparative analysis of PR-1 type proteins. Physiological and Molecular Plant Pathology *55*, 85-97.

van Loon, L.C., Rep, M., and Pieterse, C.M.J. (2006). Significance of Inducible Defense-related Proteins in Infected Plants. Annual Review of Phytopathology *44*, 135-162.

Vanderschuren, H., Lentz, E., Zainuddin, I., and Gruissem, W. (2013). Proteomics of model and crop plant species: Status, current limitations and strategic advances for crop improvement. J Proteomics *93*, 5-19.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biology *3*, 12.

Varshney, R.K., Nayak, S.N., May, G.D., and Jackson, S.A. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends in Biotechnology *27*, 522-530.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial analysis of gene expression. Science *270*, 484-487.

Verslues, P.E., Agarwal, M., Katiyar-Agarwal, S., Zhu, J., and Zhu, J.-K. (2006). Methods and concepts in quantifying resistance to drought, salt and freezing, abiotic stresses that affect plant water status. The Plant Journal *45*, 523-539.

Vertommen, A., Moller, A.L.B., Cordewener, J.H.G., Swennen, R., Panis, B., Finnie, C., America, A.H.P., and Carpentier, S.C. (2011a). A workflow for peptide-based proteomics in a poorly sequenced plant: A case study on the plasma membrane proteome of banana. J Proteomics *74*, 1218-1229.

Vertommen, A., Panis, B., Swennen, R., and Carpentier, S.C. (2011b). Challenges and solutions for the identification of membrane proteins in non-model plants. J Proteomics *74*, 1165-1181.

Vidal, R.O., do Nascimento, L.C., Mondego, J.M.C., Pereira, G.A.G., and Carazzolle, M.F. (2012). Identification of SNPs in RNA-seq data of two cultivars of Glycine max (soybean) differing in drought resistance. Genet Mol Biol *35*, 331-U258.

Vincent, D., Ergül, A., Bohlman, M.C., Tattersall, E.A.R., Tillett, R.L., Wheatley, M.D., Woolsey, R., Quilici, D.R., Joets, J., Schlauch, K.*, et al.* (2007). Proteomic analysis reveals differences between Vitis vinifera L. cv. Chardonnay and cv. Cabernet Sauvignon and their responses to water deficit and salinity. J Exp Bot *58*, 1873-1892.

Visioni, A., Tondelli, A., Francia, E., Pswarayi, A., Malosetti, M., Russell, J., Thomas, W., Waugh, R., Pecchioni, N., Romagosa, I.*, et al.* (2013). Genome-wide association mapping of frost tolerance in barley (Hordeum vulgare L.). Bmc Genomics *14*, 13.

Voets, L., Dupré de Boulois, H., Renard, L., Strullu, D.-G., and Declerck, S. (2005). Development of an autotrophic culture system for the in vitro mycorrhization of potato plantlets. Fems Microbiol Lett *248*, 111-118.

Walter, A., Scharr, H., Gilmer, F., Zierer, R., Nagel, K.A., Ernst, M., Wiese, A., Virnich, O., Christ, M.M., Uhlig, B.*, et al.* (2007). Dynamics of seedling growth acclimation towards altered light conditions can be quantified via GROWSCREEN: a setup and procedure designed for rapid optical phenotyping of different plant species. New Phytologist *174*, 447-455.

Wang, H.B., Zou, Z.R., Wang, S.S., and Gong, M. (2013a). Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L. PloS one *8*, 15.

Wang, W., Vinocur, B., Shoseyov, O., and Altman, A. (2004). Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. Trends in Plant Science *9*, 244-252.

Wang, W.X., Vinocur, B., and Altman, A. (2003). Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. Planta *218*, 1-14.

Wang, Y., Tao, X., Tang, X.M., Xiao, L., Sun, J.L., Yan, X.F., Li, D., Deng, H.Y., and Ma, X.R. (2013b). Comparative transcriptome analysis of tomato (Solanum lycopersicum) in response to exogenous abscisic acid. Bmc Genomics *14*, 14.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics *10*, 57-63.

Washburn, M.P., Wolters, D., and Yates, J.R. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nature Biotechnology *19*, 242-247.

Weckwerth, W. (2011). Green systems biology - From single genomes, proteomes and metabolomes to ecosystems research and biotechnology. J Proteomics *75*, 284-305.

Wetterstrand, K. (2014). DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). www.genome.gov/sequencingcosts (accessed on 2014/05/27)

Wienkoop, S., Morgenthal, K., Wolschin, F., Scholz, M., Selbig, J., and Weckwerth, W. (2008). Integration of metabolomic and proteomic phenotypes. Molecular & Cellular Proteomics *7*, 1725-1736.

Wienkoop, S., Baginsky, S., and Weckwerth, W. (2010). Arabidopsis thaliana as a model organism for plant proteome research. J Proteomics *73*, 2239-2248.

Wiese, S., Reidegeld, K.A., Meyer, H.E., and Warscheid, B. (2007). Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. Proteomics *7*, 340-350.

Wilkins, M.R., Gasteiger, E., Sanchez, J.C., Bairoch, A., and Hochstrasser, D.F. (1998). Two-dimensional gel electrophoresis for proteome projects: the effects of protein hydrophobicity and copy number. Electrophoresis *19*, 1501-1505.

Wisniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nat Meth *6*, 359-362.

WWAP (2012). The United Nations World Water Development Report 4: Managing Water Under Uncertainty and Risk (Paris: UNESCO).

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics *13*, 329-342.

Yang, Q.S., Wu, J.H., Li, C.Y., Wei, Y.R., Sheng, O., Hu, C.H., Kuang, R.B., Huang, Y.H., Peng, X.X., McCardle, J.A*., et al.* (2012). Quantitative proteomic analysis reveals that antioxidation mechanisms contribute to cold tolerance in plantain (Musa paradisiaca L.; ABB Group) seedlings. Mol Cell Proteomics *11*, 1853-1869.

Yazaki, J., Shimatani, Z., Hashimoto, A., Nagata, Y., Fujii, F., Kojima, K., Suzuki, K., Taya, T., Tonouchi, M., Nelson, C*., et al.* (2004). Transcriptional profiling of genes responsive to abscisic acid and gibberellin in rice: phenotyping and comparative analysis between rice and Arabidopsis. Physiological Genomics *17*, 87-100.

Yoshimura, K., Masuda, A., Kuwano, M., Yokota, A., and Akashi, K. (2008). Programmed proteome response for drought avoidance/tolerance in the root of a C(3) xerophyte (wild watermelon) under water deficits. Plant and Cell Physiology *49*, 226-241.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X*., et al.* (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science *296*, 79-92.

Yu, Y.-Q., Gilar, M., Lee, P.J., Bouvier, E.S.P., and Gebler, J.C. (2003). Enzyme-friendly, mass spectrometry-compatible surfactant for in-solution enzymatic digestion of proteins. Analytical Chemistry *75*, 6023-6028.

Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., Xie, M., Zeng, P., Yue, Z., Wang, W*., et al.* (2012). Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. Nat Biotech *30*, 549-554.

Zhang, N., Liu, B.L., Ma, C.Y., Zhang, G.D., Chang, J., Si, H.J., and Wang, D. (2014). Transcriptome characterization and sequencing-based identification of drought-responsive genes in potato. Mol Biol Rep *41*, 505-517.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PloS one *9*, e78644.

Zhu, G.Y., Geuns, J.M.C., Dussert, S., Swennen, R., and Panis, B. (2006). Change in sugar, sterol and fatty acid composition in banana meristems caused by sucrose-induced acclimation and its effects on cryopreservation. Physiol Plant *128*, 80-94.

List of publications

Articles published in internationally reviewed academic journals

- Vanhove, A., Vermaelen, W., Panis, B., Swennen, R., Carpentier, S. (2012). Screening the banana biodiversity for drought tolerance: can an in vitro growth model and proteomics be used as a tool to discover tolerant varieties and understand homeostasis. Frontiers in Plant Science, 3, doi:10.3389/fpls.2012.00176.
- Carpentier, S., Panis, B., Renaut, J., Samyn, B., Vertommen, A., Vanhove, A., Swennen, R., Sergeant, K. (2011). The use of 2D-electrophoresis and de novo sequencing to characterize inter- and intra-cultivar protein polymorphisms in an allopolyploid crop. Phytochemistry, 72, 1243-1250.

Articles published in other academic journals

- Vanhove, A., Garcia, S., Swennen, R., Panis, B., Carpentier, S. (2012). Understanding Musa drought stress physiology using an autotrophic growth system. Communications in Agricultural and Applied Biological Sciences, 77 (1), 89-93.